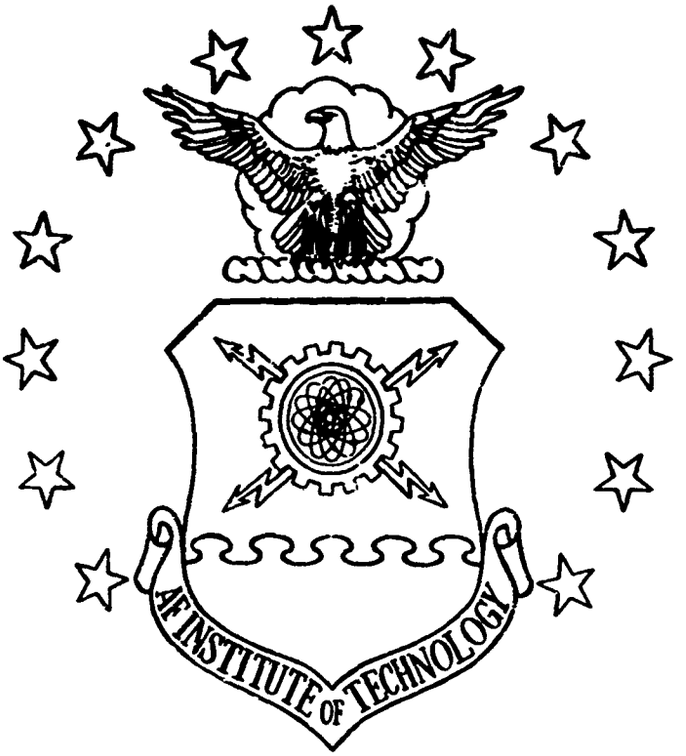


UNITED STATES AIR FORCE

AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Air Force Institute of Technology (AFIT-SE) Wright Patterson AFB, Ohio 45433	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

Robust Estimation Techniques For Location Parameter Estimation of Symmetric Distributions

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

AFIT Thesis

5. AUTHOR(S) (First name, middle initial, last name)

John Caso
Capt. USAF

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
March 1972	103	14

8a. CONTRACT OR GRANT NO. b. PROJECT NO. c. N/A d.	9a. ORIGINATOR'S REPORT NUMBER(S) GSA/MA[redacted]/72-3 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

Several robust estimators are considered for analysis and explanation. Monte Carlo techniques are used to investigate the efficiency of these robust estimators relative to the best estimator for the distribution under consideration. Sample sizes of 12 and 24 were drawn 4200 times from five symmetric probability distributions. The results show that over a class of distributions the robust estimators provide a higher guaranteed efficiency than the best estimator for any particular distribution in the family. Some interesting results are apparent from an analysis of the graphs in Appendix C indicating some upper bounds on the size of the Monte Carlo sample when conducting this type of study.

DD FORM 1 NOV 65 1473

A

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Robust Estimation						
Estimation Techniques						
Monte Carlo Techniques						
Order Statistics						

ROBUST ESTIMATION TECHNIQUES FOR LOCATION PARAMETER ESTIMATION OF SYMMETRIC DISTRIBUTIONS

THESIS

GSA/MA/72-3

JOHN CASO
CAPT. USAF

c

ROBUST ESTIMATION TECHNIQUES FOR LOCATION PARAMETER ESTIMATION OF SYMMETRIC DISTRIBUTIONS

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

by

John Caso, B.S. Mathematics
Capt. USAF

Graduate Systems Analysis

March 1972

MA
GSA/[redacted]/72-3

PREFACE

This thesis is the presentation of the results of an extensive literature search in the area of robust estimation techniques. Presently there is no formal text available on the market that gives more than an introductory look at robust estimation techniques. Most of the theory developed over this area has been presented only in statistical journals and technical reports. It was the intent of this thesis to present in a concise manner a survey of several of the most current and useful techniques.

Robust estimators were chosen which were theoretically and computationally tractable so that they could be easily understood by a practicing analyst or scientist. Section III contains a descriptive analysis of the chosen estimators and is followed by an extensive analysis of their performance using Monte Carlo techniques in Section IV.

The method of presentation assumes a basic understanding of the principles of probability and statistics. All material is presented as simply and concisely as possible. It was intended that the estimators chosen for study would be ones whose overall performance was good and which lent themselves toward application fairly easily.

GSA/MA[redacted]/72-3

I wish to thank my thesis advisor Professor A.H. Moore whose guidance contributed significantly to the completion of this study and also my thesis reader Major Ronald J. Quayle for his patience and direction.

John Caso

CONTENTS

CONTENTS

ABSTRACT

Several robust estimators were considered for analysis and explanation. Monte Carlo techniques were used to investigate the efficiency of these robust estimators relative to the best estimator for the distribution under consideration. Sample sizes of 12 and 24 were drawn 4200 times from five symmetric probability distributions. The results showed that over a class of distributions the robust estimators provided a higher guaranteed efficiency than the best estimator for any particular distribution in the family. Some interesting results are apparent from an analysis of the graphs in Appendix C indicating some upper bounds on the size of the Monte Carlo sample when conducting this type of a study.

I. BACKGROUND

Throughout many areas of scientific investigation there has been established a storehouse of information and techniques as a result of previous research and experimentation. This previous research and experimentation coupled with an ability which exists in many disciplines to isolate a situation for the purpose of observation has been an invaluable aid to the experimenter when testing a hypothesis. During this century all areas of science have turned at one time or another to mathematical statistics as an aid to scientific investigation. In the last quarter century extensive empirical investigation has given way almost completely to statistical inference and statistical testing of hypothesis. Some disciplines benefit more than others by this technique. Consider a continuum with the exact physical sciences positioned at the far left extreme and the inexact social sciences positioned at the right extreme. As you progress from left to right one notices a marked decrease in the degree of confidence that can be placed upon statistical estimates. The physicist and chemist at the far left side of the continuum have an abundance of empirically supportable evidence with which to base assumptions concerning the

distribution of the population from which they are sampling. This empirical support decreases rapidly as you move from left to right along this continuum.

Located at a point somewhere right of center on the continuum is a relatively new discipline, Systems Analysis, of which the author of this paper is a student. The basic tools of Systems Analysis are mathematics and mathematical statistics along with many of the techniques of Operations Research. A concise definitive explanation of Systems Analysis does not appear to be available and maybe not even possible. In a vague sense Systems Analysis attempts to combine pieces of information, which can be disjoint and totally unrelated, about a large or small system, and to draw inferences for basing conclusions so that a decision may be made or a course of action plotted.

Measures of central tendency e.g., mean, mode, median, are usually important statistics in all areas of investigation. Estimates of these measures are usually made based on the assumptions concerning the distribution of the sampled population. For the reasons stated earlier the sciences close to the left of the continuum have relatively little trouble in determining the form of the underlying distribution of a sample. Now consider the plight of the Systems

Analyst outlined in the following hypothetical situation.

A Department of Defense analyst is asked to estimate the total number of nuclear submarines and/or conventional submarines required to effectively defend the coastal United States from attack. Whether a point estimate or an interval estimate is required is immaterial since the same difficulties will exist in either case. There might be a large number of individual estimates which could be combined to determine the overall estimate. For example an estimate of the average speed of conventional and nuclear submarines would probably be required. It would be necessary to determine the amount of ocean area that these submarines could cover per unit of time. The natural tendency would be to take some random observations of the cruising speed of both types and then to compute the arithmetic mean. This statistic is known to be reliable when the sample is drawn from a normally distributed population. But what if this assumption was not justified. The resultant error in most cases would be small but could be catastrophically large in certain cases. Let us suppose the error was small. Consider now, however, a possible one thousand plus individual estimates that might be used

in determination of the optimal force size. How much confidence could you have in this estimate with the possibility of a small error compounded one thousand times present? In problems of this type assumptions about population distributions are difficult to make because of a usually small number of observations but even more so as a result of the uniqueness of each problem.

With this background in mind this thesis will examine some of the recent innovations in statistical estimation theory. An attempt will be made wherever possible to present the statistics considered in a manner which lends itself toward application of these statistics as opposed to a purely theoretical approach that may be of interest only to the theoretical statistician.

II. INTRODUCTION

Possibly the most important problem of statistical inference is the estimation of parameters (such as population mean, variance, etc.) from the corresponding statistics (i.e. sample mean, variance, etc.). The theory of estimation originated with problems where almost all of the statistical variability is due to measurement errors. This situation should be clearly distinguished from the opposite case where the data shows a large internal variability. It is interesting to note that Gauss introduced the normal distribution to provide an assymptotic distribution for the sample mean. That is the statistic existed before the theory for the normal distribution was developed. Throughout the years the use of the arithmetic mean has become almost sacred even though it could have easily been designed in some other form, for example omitting the three largest observations. This dogmatic use of the sample mean caused experimenters to be ignorant of the high sensitivity to deviations from normality of some of these standard procedures. In the late 1940's distribution free procedures brought relief to some of these estimation difficulties. More significant advances in this area were made throughout the 1960's. It was recognized that one

never really has a very accurate knowledge of the true underlying distribution and that the performance of some of the classical estimates is very unstable under small changes in the underlying distribution.

This paper will be confined to a study of estimation of location parameters. Throughout this paper the location parameter will denote the center of symmetry of a symmetric distribution on the real line. When the density function of a distribution is well specified there are usually several methods available to obtain large sample estimators of the location parameter. For example if f (the density function) is Uniform the the midrange is an efficient estimator of λ (the location parameter) or if f is Double Exponential (Laplace) then the median is an asymptotically efficient estimator of λ. This study will primarily be concerned with estimates of location parameters when the exact form of the underlying distribution is not known.

Statistical methods which are relatively insensitive to assumptions about their underlying distributions have been termed robust methods (Ref 3:169). This term has been extended to include estimators which have been specifically designed for estimation when the form of the underlying distribution is not known but some character of the family

of which the underlying distribution is a member is known. For example we may know that the density function is symmetric. These are called robust estimators. The main idea is that these estimators have been specifically designed for this purpose and we are not merely investigating the statistical robustness of an existing estimator for a known distribution. This study has been limited to estimators of location parameters and does not consider investigation of the estimation of scale parameters. Huber (Ref 8:93) discusses at some length the unsatisfactory aspects of attempting to estimate a scale parameter. He summarizes the reasons why this author and most statisticians have avoided this area.

> "The theory of estimating a scale parameter is less satisfactory than that of estimating a location parameter. Perhaps the main source of trouble is that there is no natural "canonical" parameter to be estimated. In the case of the location parameter, it was convienent to restrict attention to symmetric distributions; then there is a natural location parameter, namely the location of the center of symmetry, and we could seperate difficulties by optimizing the estimator for symmetric distributions (where we know what we are estimating) and then investigate the properties of this optimal estimator for non standard conditions, e.g., for nonsymmetric distributions. In the case of the scale parameter, we meet, typically, highly symmetrical distributions, and the above device to ensure unicity of the parameter to be estimated fails. Moreover it becomes questionable, whether one should minimize bias or variance of the estimator.
> So we shall just go ahead and shall construct estimators that are invariant under s[illegible] transformations and

that estimate their own asymptotic values as accurately as possible. Of course one has to check afterward in a few typical cases what these estimators really do estimate".

Specifically this thesis has two purposes.

1. To provide a survey of the current techniques involved in robust estimation of a location parameter of a symmetric probability distribution.
2. To explore, using Monte Carlo techniques, the performance of some selected robust estimators of a location parameter. The purpose of this investigation will be to ascertain what benefit, if any, can be achieved through the use of robust estimators making no assumptions about the specific form of the underlying probability distribution, as opposed to employing known estimators for a predetermined probability distribution.

To achieve these objectives an extensive literature search and study was made of the available literature. The results are presented in this thesis. The bibliography contains a listing of those sources found to contain much of the applicable information on robust estimation which were used directly in the formulation of this paper. Appendix A

is a supplemental bibliography which contains a listing of those sources which were either applicable to robust estimation techniques and were not available, which apply only to statistical robustness in general, or were sources of a general nature which were useful in the formulation of this paper.

III. TYPES OF ROBUST ESTIMATORS

The purpose of this section is to present in a summarized form several robust estimators which have been developed for the purpose of estimating the center of symmetry of an unspecified distribution. The estimators considered will be of two basic types. One type has the characteristic that the functional form of the estimator does not depend on the sample while the other type has the actual functional form of the estimator determined by the information contained in the sample.

Pioneer efforts in robust estimation were mainly concerned with departures from the assumptions of normality. By appealing to the central limit theorem many distributions can be considered to be approximately normal. However it is easily demonstrated that even a slight departure from the assumption of normality can often cause the sample mean to behave badly as an estimator of the location parameter. Early studies were devoted primarily to estimation methods where the underlying distribution was the standard normal but was contaminated in some manner by a distribution, usually normal, with a larger amount of dispersion. More recent inquiries consider situations involving more

varied sets of distributions. In most cases only symmetric unimodal distributions are considered. Tukey (Ref 13:30) emphasizes this restriction when he considers the treatment of "spotty data".

> "Accordingly it will be for us to begin with long tailed distributions which offer the minimum of doubt as to what should be taken as the true value. If we stick to symmetric distributions we can avoid all difficulties of this sort..... No other point on a symmetrical distribution has a particular claim to be considered the true value. Thus we will do well by restricting ourselves to symmetric distributions".

This quote is presented here because throughout all the more recent papers dealing with robust estimation the restriction of a symmetrical distribution appears to have been strictly adhered to and the Tukey (Ref 13:1) paper given as the reference source. This quote will also provide some justification for the structure of the Monte Carlo analysis presented in the next section of this thesis.

The first robust estimators considered here will be of the type where the specific form of the estimator does not depend on the information contained in the sample. The functional form of these estimators will be structured as order statistics.

This type of estimator was analyzed rather extensively by Croe and Siddiqui. A class, F, of distributions are considered which are normal, cauchy, parabolic, triangular, and rectangular. The results presented claim that asymptotic efficiencies of at least .82 relative to the best estimator for a single distribution are achieved by the best trimmed mean or linearly weighted mean (Ref 3:353).

Estimators Which are Special Symmetrical Linear Combinations of Order Statistics

Two estimators of this type will be considered here. The winsorized mean and the trimmed mean. These estimators have been present for many years but were used very sparingly. The theoretical basis for winsorized and trimmed means is the technique called rejection of outliers. The trimming removes equal numbers of the highest and lowest observations and then proceeds with the remainder as if it were a complete sample. If the samples do come from a normal distribution there will be some loss in efficiency and there will be an increase in efficiency when the samples are from a distribution with long tails.

Let $X_1 \leq X_2 \leq \cdots \leq X_n$ be the order statistics resulting from random sampling of $F(X-\lambda) \in G$

and subsequent ordering. G consists of distributions which are symmetric about the median λ and such that λ is the unique mode. The amount of trimming or winsorizing p is determined here such that

$$p = 1/2 - r/n \qquad (3\text{-}1)$$

where r is a non-negative integer less than $n/2$.

Winsorized Mean: (Ref 13:1).

$$W_n(p) = n^{-1}\left[(r+1)\left(X_{r+1} + X_{n-r}\right) + \sum_{i=r+2}^{n-r-1} X_i\right] \qquad (3\text{-}2)$$

and if $n = 2v+1$

$$W_n(1/(2n)) = X_{v+1} \qquad (3\text{-}3)$$

where $r < (n-1)/2$

Trimmed Means (Ref 13:1).

$$T_n(p) = (n-2r)^{-1} \sum_{i=r+1}^{n-r} X_i \qquad (3\text{-}4)$$

In a normal sample winsorized means are more stable than trimmed means. The possible loss in efficiency through the use of these estimators is far overshadowed by the large possible gain when the assumption of normality is violated. Many papers published in the area of robust estimation have dealt with linear combinations of order statistics (Ref 1, 3, 4, 5) and in many cases much of the analysis centered around winsorized and trimmed means. For an in depth discussion of this area see the paper by Gastwirth and Rubin (Ref 5). Gastwirth and Rubin demonstrate that within a large class of estimators there is a unique maximum efficient linear estimator. The difficulty of determining a maximum efficiency linear estimator for specific families of densities is emphasized and the paper is restricted to searching for maximum efficient estimators in smaller classes of linear estimators such as the trimmed means and linear combinations of a few sample percentiles.

Estimators Which are not Strictly Functions of Order Statistics

This section will develop the estimators in the same manner as the previous section. While many of these estimators, which make up the greater part of this study, do utilize order statistics they are not considered strictly functions of order statistics as were the estimators in the previous section. These estimators were chosen from the many which exist today for several reasons. First of all they have been shown in some previous studies to possess a high relative efficiency over fairly broad classes of distributions. Secondly they are, in most cases, theoretically simple to comprehend and computationally tractable to apply.

Hodges-Lehmann Estimator (Ref 6).

This estimator was one of the earlier attempts at the development of a robust estimator and from most results in the literature appears to be one of the best estimators. Results obtained by Bickel (Ref 1) indicate that, in terms of robustness, the Hodges-Lehmann estimate is superior to the trimmed and winsorized means. It is simple in form and computationally easy to handle. Hodges (Ref 6) defined this estimator of the location parameter in terms

of rank test statistics such as the Wilcoxan or Normal scores statistic.

Let $X_1, X_2, \ldots\ldots, X_n$ be a random sample from an unknown symmetric probability distribution. Then

$$HL\left[X_1, X_2, \ldots, X_n\right] = \underset{i \leq j}{MED}\left[\frac{X_i + X_j}{2}\right] \qquad (3\text{-}5)$$

$$i, j = 1, 2, \ldots, n$$

This estimator is formed by taking the median of the mean of all of the $\binom{n}{2}$ pairs in the sample. In the Hodges-Lehmann paper (Ref 6) it is shown that the estimates are symmetric with respect to the parameter being estimated and thus to be unbiased if the underlying distribution of the observations on which the estimate is based is symmetric. The form of the estimator makes it the only practically tractable estimator derived from the ranks test. When the sample gets very large, however, the number of steps involved becomes prohibitive. An alternate form of this estimator which uses ordered samples and is much quicker to compute for large samples has shown to be good in certain situations

Let $X_1, X_2, \ldots, X_n$ be an ordered random sample.

$$HL2[X_1, X_2, \ldots, X_n] = MED\left[\frac{X_i + X_{n+1-i}}{2}\right] \quad (3\text{-}6)$$

$$1 \leq i \leq \frac{n+1}{2}$$

Huber's Estimator (Ref 8).

Huber deals with the asymptotic theory of estimating a location parameter. The emphasis in this paper was placed on treating contaminated normal distributions. There seems to be some discussion over just how well this estimator actually performs. It is presented here in summary form mainly because Huber does attempt to design a robust estimator of the scale parameter. Basically Huber considers the method of least squares where the idea is to minimize the expression

$$\sum_i (X_i - \xi)^2 \quad (3\text{-}7)$$

where $X_1, X_2 \ldots, X_n$ is a random sample and

$$\xi = \sum_i X_i/n \tag{3-8}$$

Huber's approach was to search for a $\tilde{\xi}$ such that

$$\tilde{\xi} = \tilde{\xi}_n(X_1, X_2, \ldots, X_n) \tag{3-9}$$

minimizes

$$\sum_i \rho(X_i - \tilde{\xi}) \tag{3-10}$$

where

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & |t| < k \\ k|t| - \frac{1}{2}k^2 & |t| \geq k \end{cases} \tag{3-11}$$

$$t_i = |X_i - \tilde{\xi}|$$

k was here related to the contamination proportion. Huber showed that taking $k = 2.$ will do well for any contamination proportion less than twenty percent. Huber obtained robust estimators T and S for both the location and scale parameters as the solutions of the following simultaneous equations.

$$n^{-1}\sum_{i=1}^{n} \omega\left(k, \frac{X_i - T}{S}\right) = 0 \qquad (3\text{-}12)$$

$$n^{-1}\sum_{i=1}^{n} \omega^2\left(k, \frac{X_i - T}{S}\right) = E_\eta\, \omega^2(k, X) \qquad (3\text{-}13)$$

where

$$\omega(k, X) = \begin{cases} X & |X| < k \\ k\ \operatorname{sgn} X & |X| \geq k \end{cases} \qquad (3\text{-}14)$$

and

$$E_\eta\, \omega^2(k, X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \omega^2(k, X)\, e^{-\frac{X}{2}}\, dx \qquad (3\text{-}15)$$

Solutions to these equations for T and S along with an iterative computational procedure can be found in Leone (Ref 10). Leone performed a Monte Carlo study using contaminated normal distributions drawing a sample of size 20. The sample was drawn 500 times. The results indicate that the variance of the estimators increased with an increase in the scale parameter of the contaminating distribution but are less sensitive to a change in the location parameter.

The remaining estimators to be considered in this section are the type whose functional form is determined by the information contained in the sample. Takeuchi (Ref 12:292) called this type quasilinear estimators.

Quasilinear Estimators

The estimators of this type considered here will normally use some known statistic for information with which to base a choice between several competing functional forms available. The various functional forms used in the following estimators were chosen on the basis of some very weak assumptions about the general form that the underlying distribution might possess. Thus it should be clear that

these choices are merely examples and the intent of the analysis here is to emphasize the great flexability available in choosing these competing functional forms, making these very powerful estimators.

Keeping the notation consistent with that used previously G will denote a family of distributions which are symmetric about the median λ and such that λ is the unique mode.

Switzer's Estimator (Ref 11).

The method employed here is to choose from a set of competing estimators that estimator which has the minimum standard error for the sample being considered. The forms of the competing estimators are predetermined but which one is chosen is determined by the information contained in the sample. In formulating this estimator Switzer outlines two fairly loose restrictions on the set of competing estimators from which to choose:

1. that the competing estimators be such that their standard errors can also be estimated without making use of the unknown shape of the underlying distribution and,
2. that the collection should contain only estimators whose efficiency relative to one another ranges from

very small to very large numbers as the distributions range over a set of reasonable possibilities.

Estimators are chosen in the manner described for each available sample of equal size and a sequence of estimates of the location parameter is obtained. Switzer chose to limit the number of competing estimators to three. This paper's author will continue this convention throughout this thesis. However it is apparent that this principle could be extended to include a larger number of competing estimators and is so suggested by Switzer at the close of his paper.

Let ξ_i $i=1,2,3$ be three sequences of competing estimates obtained from three selected estimators which are defined for every sample size N and let S_i $i=1,2,3$ be non-parametric estimates of the standard errors. Then the recommended estimator is:

$$SW = \sum_{i=1}^{3} Z_i \xi_i \tag{3-16}$$

$$Z_i = \begin{cases} 1 \text{ if } \min[S_1, S_2, S_3] = S_i \\ 0 \text{ otherwise} \end{cases}$$

It is assumed here that $\sqrt{N}\left|\xi_i - \lambda\right|$ has a limiting normal distribution with 0 mean for $i=1,2,3$ and that the standard error estimates were chosen so that NS_i^2 consistently estimates $\sigma_i^2(f)$ for each i, and f belonging to a large class G. If $\xi(f)$ is the most efficient of the three competitors for a given f then $\sqrt{N}\left|SW - \lambda\right|$ has the same limiting distribution as $\sqrt{N}\left|\xi(f) - \lambda\right|$ for all $f \in G$.

Switzer outlines two general procedures for obtaining non parametric estimates of the standard errors of the competing estimators. Only one procedure will be presented here. It is a two step procedure.

Step 1. Assume the sample can be divided into K blocks of equal size $n = N/K$.

Step 2. Compute ξ_i based on samples of size n.

$$\xi_i^k \quad k=1,2,\ldots K \qquad i=1,2,3$$

$$\xi_i = \sum_1^k \xi_i^k / K \tag{3-17}$$

and

$$S_i^2 = \sum_1^k \left| \xi_i^k - \xi_i \right|^2 \Big/ K|K-1| \qquad (3\text{-}18)$$

Specifically, Switzer chose for study an N (sample size) which was divisible by six and computed the three mid ranges.

$$\xi_1^k = \left[X_{|3|} + X_{|4|} \right] \Big/ 2 \qquad (3\text{-}19)$$

$$\xi_2^k = \left[X_{|2|} + X_{|5|} \right] \Big/ 2 \qquad (3\text{-}20)$$

$$\xi_3^k = \left[X_{|1|} + X_{|6|} \right] \Big/ 2 \qquad (3\text{-}21)$$

$k = 1, 2, \ldots, K$

The results of the Monte Carlo analysis performed by Switzer using his estimator as previously outlined showed that the SW estimator performed very well when the sample was drawn from short, long, and normal tailed distributions with samples of size 30, 60, and 120. In each case the SW estimator was not quite as good as the best estimator (sample mean, median, mid range, etc.) for that particular shape. It was shown however, to always have less variance than the other estimators considered for that shape.

Hogg's Estimator (Ref 7).

This estimator is very interesting because of the large amount of possibilities it presents and very appealing because of its extreme simplicity. Hogg uses the kurtosis of the sample to determine which form the estimator should take. Kurtosis here being defined as the fourth central moment divided by the square of the variance. The sample kurtosis

$$k = \sum_{1}^{n} \frac{\left[X_i - \bar{X}\right]^4}{n S^4} \qquad (3\text{-}22)$$

where n is the sample size and $\bar{X}$ is the sample mean, converges in probability to the kurtosis of the underlying distribution of the sample. Hogg subsequently

structured his estimator in the following manner.

$$HG = \begin{cases} \bar{X}^{C}_{1/4} & k < 2. \\ \bar{X} & 2. \leq k \leq 4. \\ \bar{X}_{1/4} & 4. < k \leq 5.5 \\ M & 5.5 < k \end{cases} \qquad (3\text{-}23)$$

where

$\bar{X}^{C}_{1/4}$ is the mean of the $n/4$ smallest and $n/4$ largest items of the sample.

$\bar{X}_{1/4}$ is the mean of the remaining interior sample items.

$\bar{X}$ is the sample mean.

M is the sample median.

The many possibilities of this estimator should now be apparent to the reader for there is really no restriction on the possible ranges of k or the choice of forms for the estimator. Based solely on the kurtosis of the sample this estimator might prove useful indeed if its corresponding

results were fruitful. Hogg performed a Monte Carlo analysis in which the performance of his estimator was compared mainly with the performance of the Hodges-Lehmann estimator. The analysis was performed over a class of distributions ranging from Rectangular to Cauchy. Hogg's estimator performed better overall than the Hodges-Lehmann estimator which also performed very well.

It is possible to generalize Hogg's estimator in such a manner that the estimator is a linear combination of the sample items with weights which are continuous functions of the sample items. The procedure is summarized below. See Hogg (Ref 7:1184) for a more complete discussion.

If $X_1, X_2, \ldots, X_n$ are sample values then

$$HG2 = \sum_{i=1}^{n} W_i \cdot X_i \qquad (3\text{-}24)$$

where

$$W_i = {}^{1}/_{V_i} \Big/ \sum {}^{1}/_{V_i} \qquad (3\text{-}25)$$

and

$$V_i = \max\left[1+\frac{.03}{S^2}\left|k-3\right|\left|X_i-M\right|^3,\ 0.01\right] \quad (3\text{-}26)$$

The author notes that if $k>3$, this statistic places less weight on the extreme observations and with $k<3$ it assigns more weight to the extreme observations.

IV. MONTE CARLO ANALYSIS

This investigation was conducted to explore the performance of three of the estimators discussed in the previous section. The three chosen were the Hodges-Lehmann estimator, the Switzer estimator, and Hogg's estimator. There were two reasons for selecting these three estimators from the many which can be found in the available literature. First of all the Hodges-Lehmann estimator and Hogg's estimator had demonstrated a high degree of efficiency in estimating location parameters in much of the analysis found in the literature. The Switzer estimator is very new and could be found in only one article (Ref 11). The Switzer estimator does however demonstrate a new and interesting technique for exploration. Thus in an attempt to test the performance of the Switzer estimator it was necessary to select what are generally considered the best available robust estimators as competitors. The second reason was that these estimators had not previously been compared against one another for these sample sizes and probability distributions and also that these, as with all the estimators considered in the previous section, were computationally and theoretically manageable.

The analysis was basically a computer exercise and all computations were performed on the Control Data Corporation 6600 Computer System. Five basic probability distributions were selected which were symmetric and unimodal. Utilizing Monte Carlo techniques, random samples of size 12 and 24 were drawn from these five distributions. At the outset of the analysis several larger sample sizes were drawn but the additional gain in information did not prove to be worth the extra cost in computer time so these larger sample sizes were eliminated. Using the random samples drawn, estimates of the location parameter were computed using the robust estimators and also using the known statistics which are the "best" estimators for each of the distributions considered. The final step was to compute the variance from the true value of the location parameter for each of the estimates. The computer program listing of the program designed to accomplish this procedure can be found in Appendix B.

Probability Distributions Used

Samples of size 12 and 24 were drawn from each of five probability distributions. As stated earlier each was a symmetric distribution. The specific distributions were

selected because of their similarity in the sense that they could easily be mistaken for one another when a decision maker had to base a decision on a small sample. It is also possible for these distributions to occur in combination with one another thus causing further confusion.

Rectangular.

$$F[X] = 1 \qquad 0 \leq x \leq 1 \qquad (4\text{-}1)$$

Triangular.

$$F1[X] = [2/(a+b)a]\cdot[a+x], \quad -a \leq x \leq 0 \qquad (4\text{-}2)$$

$$F2[X] = [2/(a+b)b]\cdot[b-x], \quad 0 \leq x \leq b \qquad (4\text{-}3)$$

Drawings were made from this distribution with three different parameters.

$$-1 \leq x \leq 1$$

$$-5 \leq x \leq 5$$

$$-10 \leq x \leq 10$$

Normal.

$$F[x] = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \qquad (4\text{-}4)$$

Contaminated Normal.

10% Contamination

$$F[x] = .90\left[\frac{1}{\sqrt{2\pi}}\right]\exp\frac{-x^2}{2} + .10\left[\frac{1}{\sqrt{6\pi}}\right]\exp\frac{-x^2}{6} \qquad (4\text{-}5)$$

20% Contamination

$$F[x] = .80\left[\frac{1}{\sqrt{2\pi}}\right]\exp\frac{-x^2}{2} + .20\left[\frac{1}{\sqrt{6\pi}}\right]\exp\frac{-x^2}{6} \qquad (4\text{-}6)$$

Double Exponential.

$$F[x] = \tfrac{1}{2}\exp|x| \qquad (4\text{-}7)$$

Estimators Considered

The three robust estimators analyzed were used in the forms stated earlier in this text, i.e., equations 3-5, 3-16, and 3-23. Three popular statistics were also computed, the sample mean, the sample median, and the mid range. The form of these statistics should be familiar to anyone with an interest in statistics.

Computations

The variance of each estimator with respect to the true value of the location parameter was computed using the mean square error. Each sample size was drawn 4200 times.

$$VAR = \sum_{i=1}^{4200} \frac{[\hat{\lambda}_i - \lambda]^2}{4200} \qquad (4\text{-}8)$$

The computer program was designed to compute the mean square error every 350 repetitions and provide these values as outputs. Appendix C contains graphs of some selected results obtained for some of the estimators from each distribution considered. The majority of the graphs were omitted from this thesis to keep the size manageable

and it was the opinion of this author that they would not provide any meaningful information.

Relative efficiencies were also computed and the results are expressed in per-centage form and presented in Appendix D. Relative efficiency is defined here to be the ratio of the variance of the best estimator for the distribution considered to the estimator whose efficiency is under consideration.

V. CONCLUSIONS

Since this thesis was intended mainly to provide a survey of existing robust estimation techniques the content is not easily extrapolated to many significant conclusions. Several interesting observations however, were made in the course of this study which are worthy of note.

First of all the technique of robust estimation as it has been presented here is not over 10 years old. The scarcity of practical estimation techniques and an absence of a theoretical foundation for this discipline emphasizes its newness. The fact that investigation in this broad and interesting area of statistics has h rdly scratched the surface is a conclusion worthy of mention. There appears to be approximately ten theoretical statisticians who are doing the majority of the research in this area. Their names can be found in the bibliographies at the end of this thesis. The amount of duplicated effort evident in the literature is testimony to the infancy of this discipline.

Several interesting conclusions can be made based on the results of the Monte Carlo analysis summarized in Tables I thru VI in Appendix D. As stated earlier these

estimators were designed to estimate the location parameter of a symmetric distribution. The relative efficiencies presented in Appendix D denote the performance of each estimator relative to the "best" estimator for that distribution. Obviously if the exact form of the underlying distribution was known the "best" estimator could easily be selected. Suppose however, that a sample of size 12 was drawn from either a Normal, Contaminated Normal, or Double Exponential distribution, with equal probability. Tables I and V show that if the Sample Mean were selected to estimate the location parameter the highest efficiency that could be achieved would be 100% and the lowest efficiency would be 72.7%. If however, the Hodges-Lehmann estimator was chosen the highest efficiency would be 100% (efficiencies in the Tables greater than 100% are taken here as 100%) and the lowest 92.1%. If Hogg's estimator were chosen the high would have been 98.5% and the low 78%. Now consider all five distributions from Tables I thru VI. Suppose the Mid Range was selected as the estimator of the location parameter. Then the efficiency would range from 100% to 6.2%. Once again if the Hodges-Lehmann estimator was chosen the efficiency would range from 100% to 32.5%. The Hodges-Lehmann estimator is truly

superior to the Mid Range when the efficiencies are compared for each distribution. The data presented in Appendix D shows that for the estimators and distributions considered the robust estimators are superior.

Several minor conclusions are worthy of mention here. First of all a comparison of the efficiencies for sample sizes 12 and 24 show that as the sample size gets larger the "best" estimator gets better and the efficiency of the robust estimator decreases. This is to be expected however, since the robust estimators were designed and have value only for small sample sizes. Another conclusion of some import is that varying the scale parameter of the underlying distribution has no effect on estimation of the location parameter. This is evident from Tables I, II, III, and IV.

The final conclusion has more application in the area of Monte Carlo techniques than robust estimation techniques. During the course of this investigation there was some question as to the number of times each sample should be drawn. The low figure was approximately 500 drawings and the high figure 5000. The graphs presented in Appendix C show that for all practical purposes 1000 repetitions would be sufficient and that any over 2000 is just not worth the computer time.

Areas For Further Investigation

Initially the author of this paper felt that the Switzer estimator would have the most efficient form. The choice of competing forms of the estimator did not bear out this premise as was demonstrated in the analysis. If however, a more judicious choice of competing estimators was made the performance of this estimator might be significantly enhanced. This area plus the possibility of analyzing the performance of these robust estimators when the restriction of a symmetric underlying distribution does not apply could be extremely fruitful areas for further study.

Bibliography

1. Bickel, P.J., On Some Robust Estimates Of Location. Ann. Math. Statist. 36:847-858(1965)

2. Birnbaum, A. & E. Laska. Optimal Robustness: A General Method With Applications To Linear Estimators Of Location. J. Amer. Statist. Assoc. 62:1230-1240 (1967)

3. Crow, E. L. & M. Siddiqui. Robust Estimation Of Location. J. Amer. Statist. Assoc. 62:353-388(1967)

4. Gastwirth, J. L. On Robust Procedures. J. Amer. Statist. Assoc. 61:929-948(1966)

5. ---------- & H. Rubin. On Robust Linear Estimators. Ann. Math. Statist. 40:24-39(1969)

6. Hodges, J. L., Jr. & E. L. Lehmann. Estimates Of Location Based On Rank Tests. Ann. Math. Statist. 34:598-611(1963)

7. Hogg, R. V. Some Observations On Robust Estimation. J. Amer. Statist. Assoc. 62:1179-1186(1967)

8. Huber, P. J. Robust Estimation Of A Location Parameter. Ann. Math. Statist. 35:73-102(1964)

9. Lehmann, E. L. Robust Estimation In Analysis Of Variance. Ann. Math. Statist. 34:957-66(1963)

10. Leone, Fred C., Toke Jayachandran, & Stanley Eisenstat. A Study Of Robust Estimators. Tech. 9:652-660(1967)

11. Switzer, Paul. Comments And Suggestions On Efficiency Robustness. Technical Report no. 163. Dept. Of Statistics, Stanford University.

12. Takeuchi, K. Uniformly Asymptotic Efficient Estimator Of A Location Parameter. J. Amer. Statist. Assoc. 66:292-301(1971)

13. Tukey, John W. The Future Of Data Analysis. Ann. Math. Statist. 33:1-67(1962)

14. Van Eeden, C. Efficiency Robust Estimation Of Location. Ann. Math. Statist. 41:172-181(1970)

APPENDIX A

SUPPLEMENTAL BIBLIOGRAPHY

1. Allen, David P. The Robustness Of The Students T Test When Sampling From A Weibull Distribution. Masters Thesis. Naval Postgraduate School. Monterrey, Calif.

2. Anscombe, F.J. Rejection Of Outliers. Tech. 2:123-7 (1960)

3. Bloch, Daniel A. & J.L. Gastwirth. On Asymptotically Robust Competitors Of The One Sample T Test. Technical Report no. 70. Dept. Of Statistics. Johns Hopkins University. Baltimore, Maryland.

4. Birnbaum, A. Some Theory And Techniques For Robust Estimation. (Abstract) Ann. Math. Statist. 32:622(1961)

5. ---------- & E. Laska. Efficiency Robust Two Sample Rank Tests. J. Amer. Statist. Assoc. 62:1241-51(1967)

6. ---------- Optimal Robustness For Estimators And Tests. Technical Report. Courant Institute Of Mathematical Sciences. New York University

7. Birnbaum, A. & Mike, V. Asymptotically Robust Estimators Of Location. J. Amer. Statist. Assoc. 65:1265-82(1970)

8. Birnbaum, A., E. Laska & M. Meisner. Optimally Robust Linear Estimators Of Location. J. Amer. Statist. Assoc. 66:302-10(1971)

9. Box, G.E.P. & G.C. Tiao. A Further Look At Robustness via Bayes' Theorem. Biometrica 51:169-73 (1964)

10. ---------- A Note On Criterion Robustness And Inference Robustness. Biometrica 49:419-32(1962)

11. Craig, Allen T. & Robert V. Hogg. Introduction To Mathematical Statistics. New York: Macmillan Co., 1970

12. Filliben, J.J. Simple And Robust Linear Estimator Of The Location Parameter Of A Symmetric Distribution. Doctoral Dissertation. Princeton University. (1969)

13. Gastwirth, J.L. "On Robust Rank Tests" in Non Parametric Techniques In Statistical Inference, edited by Madav Lal Puri. Cambridge At The University Press. 1970.

14. ----------& M.S. Cohen. Small Sample Behavior Of Robust Linear Estimators Of Location. J. Amer. Statist. Assoc. 65:946-73(1970)

15. ----------& H. Rubin. The Behavior Of Robust Estimators On Dependent Data. Technical Report. Dept. Of Statistics. Purdue University. Layfayette, Ind.

16. Govindarajulu, Zakkula. Certain General Properties Of Unbiased Estimates Of Location And Scale Parameters Based On Ordered Observations. Siam Journal Of Applied Mathematics. 16:533-51(1968)

17. Hampel, F.R. Contributions To The Theory Of Robust Estimation. Masters Thesis. Berkeley, Calif.

18. Hatch, L.O. & H. Posten. A Quantitative Approach To Robustness. Research Report no. 42. University Of Connecticut. Storrs, Connecticut. 1968

19. ----------Robustness Of The Student Procedure. Research Report no. 24. University Of Connecticut. Storrs, Connecticut. 1966

20. Hillier, Fredrick S. & Gerald S. Lieberman. Introduction To Operations Research. San Francisco: Holden-Day Inc. 1970

21. Huber, P.J. Robust Estimation. Mathematical Centre Tracts. 27:3-25(1968)

22. ----------"Studentizing Robust Estimates" in Non Parametric Techniques In Statistical Inference, edited by Madav Lal Puri. Cambridge At The University Press. 1970.

23. Jaeckel, Louis A. Some Flexable Estimates Of Location. Ann. Math. Statist. 42:1540-52(1971)

24. ---------- Robust Estimates Of Location. Thesis. Berkeley, Calif.(1969)

25. Kendall, M.G. & A. Stuart. The Advanced Theory Of Statistics Vols. I, II, III. New York. Hafner Publishing Co., 1961.

26. Laska, E. A General Theory Of Robustness. Doctoral Dissertation. New York University, 1962.

27. Lehmann, E.L. & J.L. Hodges Jr. Basic Concepts Of Probability And Statistics. San Francisco: Holden-Day Inc., 1970.

28. Mike, V. Contributions To Robust Estimation. Technical Report no. 361. Courant Institute Of Mathematical Sciences. New York University, 1967.

29. --------- Robust Systematic Linear Estimators Of Location. J. Amer. Statist. Assoc. 66:594-601(1971)

30. Moy, William A. "Monte Carlo Techniques: Practical". In The Design Of Computer Simulation Experiments, edited by Thomas H. Naylor. Durham N.C.: Duke University Press, 1969.

31. Pratt, J.W. Robustness Of Some Procedures For The Two Sample Location Problem. J. Amer. Statist. Assoc. 59:665-80(1964)

32. Sen, Pranab Kumar. Robustness Of Some Non Parametric Procedures In Linear Models. Ann. Math. Statist. 39:1913-22(1968)

33. Scheffe, Henry. The Analysis Of Variance. New York: Wiley & Sons Inc., 1959.

34. Siddiqui, M. & K. Raghunandanan. Asymptotically Robust Estimators Of Location. J. Amer. Statist. Assoc. 62:950-53(1967)

35. Yhap, E. F. An Asymptotically Optimally Robust Linear Unbiased Estimator Of Location For Symmetric Shapes. Doctoral Dissertation. New York University, 1967.

36. ---------- Asymptotic Optimally Robust Linear Unbiased Estimators Of Location And Scale Parameters. To be submitted to Ann. Math. Statist.

APPENDIX B

COMPUTER PROGRAM LISTING

```
      PROGRAM MAIN(INPUT,OUTPUT,PLOT)
      DIMENSION X(20)
      DIMENSION Y(20)
      DIMENSION Z(20)
      CALL PLOT(1.,2.,-3)
      READ 500,N
      DO 5 JJ=1,4
      DO 3 II=1,6
      DO 1 I=1,N
    1 READ 501,X(I)
      TEMP=0.
      DO 2 J=1,N
      TEMP=X(J)+TEMP
      Y(J)=TEMP/J
      GO TO(10,11,12,13,14,15),II
    4 M=J*350
      Z(J)=J
    2 PRINT 502,M,Y(J)
      CALL GRAPH(Z,Y,N)
    3 CONTINUE
    5 CONTINUE
      GO TO 7
   10 PRINT 601
      GO TO 4
   11 PRINT 602
      GO TO 4
   12 PRINT 603
      GO TO 4
   13 PRINT 604
      GO TO 4
   14 PRINT 605
      GO TO 4
   15 PRINT 606
      GO TO 4
    7 CONTINUE
  601 FORMAT(28X*HODGES LEHMANN ESTIMATOR*)
  602 FORMAT(28X*HOGGS ESTIMATOR*)
  603 FORMAT(28X*SWITZERS ESTIMATOR*)
  604 FORMAT(28X*SAMPLE MEAN*)
  605 FORMAT(28X*SAMPLE MEDIAN*)
  606 FORMAT(28X*MID RANGE*)
  500 FORMAT(I3)
  501 FORMAT(F14.9)
  502 FORMAT(//,28X*AFTER *I4* REPETITIONS*//,28X,F14.10)
      END
```

```
      PROGRAM MAIN(INPUT,OUTPUT,PUNCH)
      DIMENSION RAND(48),RORN(48),CONT(48),TRIAN(48),EXP(48)
      DIMENSION VR(50),VC(50),VT(50),VN(50)
      DIMENSION VE(50)
      READ 600,AVG,PCT,STD,A,N
      READ 501,S
      CALL RANSET(S)
      X=S
      PRINT 600,X
      DO 50 JA=1,12
      PRINT 601,AVG,PCT,STD,A,N
      DO 18 JJ=12,24,12
      VR(JJ)=0.
      VC(JJ)=0.
      VT(JJ)=0.
      VE(JJ)=0.
      VN(JJ)=0.
   18 CONTINUE
      DO 1 JI=1,N
      DO 1 I=12,24,12
      Y=JI
      K=I
      DO 2 J=1,K
      X=RANF(Y)
    2 RAND(J)=X
      CALL GAUSS(K,JI,RORM)
      CALL CNORM(STD,PCT,K,AVG,RORM,CONT,JI)
      CALL TRIANG(A,K,TRIAN,JI)
      CALL EXPON(K,EXP,JI)
      CALL AHODG(RAND,K,ARORD)
      CALL AHODG(CONT,K,ACORD)
      CALL AHODG(TRIAN,K,ATORD)
      CALL AHODG(RORM,K,ARORD)
      CALL AHODG(EXP,K,AEXPO)
      VR(K)=((ARO-0.5)**2)+VR(K)
      VC(K)=(ACORD**2)+VC(K)
      VT(K)=(ATORD**2)+VT(K)
      VN(K)=(ARORD**2)+VN(K)
      VE(K)=(AEXPO**2)+VE(K)
    1 CONTINUE
      DO 20 KK=12,24,12
      VR(KK)=VR(KK)/N
      VC(KK)=VC(KK)/N
      VT(KK)=VT(KK)/N
      VN(KK)=VN(KK)/N
      VE(KK)=VE(KK)/N
      PRINT 512,N,KK,VR(KK),VC(KK),VT(KK),VN(KK),VE(KK)
      PUNCH 501,VR(KK),VC(KK),VT(KK),VN(KK),VE(KK)
   20 CONTINUE
   50 CONTINUE
      CALL RANGET(Y)
      PUNCH 501,Y
  501 FORMAT(F14.9)
  512 FORMAT(23X*VARIANCE OF ESTIMATES(MEAN SQUARE ERROR) AFTER *I3,/,23
     $X*REPETITIONS WITH SAMPLE SIZE=*I2,//,23X*HODGES LEHMANN*//,23X,5F
     $14.9)
  600 FORMAT(1X,F15.9)
  601 FORMAT(2X,4F15.9,I3)
  500 FORMAT(F5.2,F5.2,F8.3,F5.2,I3)
      END
```

```
      PROGRAM MAIN(INPUT,OUTPUT,PUNCH)
      DIMENSION XR(50),XC(50),XT(50),XN(50),XE(50)
      DIMENSION SR(50),SC(50),ST(50),SN(50),SE(50)
      DIMENSION RAND(48),RORM(48),CONT(48),TRIAN(48),RORD(48)
      DIMENSION HR(50),HC(50),HT(50),HN(50),HE(50)
      DIMENSION EXP(48),EXPOD(48)
      DIMENSION ROMD(48),CORD(48),TORD(48)
      DIMENSION STATR(8),STATC(8),STATT(8),STATM(8),STATE(8)
      DIMENSION RR(50),RC(50),RT(50),RN(50),RE(50)
      DIMENSION YR(50),YC(50),YT(50),YN(50),YE(50)
      READ 500,AVG,PCT,STD,A,N
      READ 501,S
      CALL RANSET(S)
      X=S
      PRINT 600,X
      PRINT 601,AVG,PCT,STD,A,N
      DO 18 JJ=12,24,12
      SE(JJ)=0.
      SN(JJ)=0.
      ST(JJ)=0.
      SC(JJ)=0.
      SR(JJ)=0.
      YR(JJ)=0.
      YC(JJ)=0.
      YT(JJ)=0.
      YN(JJ)=0.
      YE(JJ)=0.
      XR(JJ)=0.
      XC(JJ)=0.
      XT(JJ)=0.
      XN(JJ)=0.
      XE(JJ)=0.
      HR(JJ)=0.
      HC(JJ)=0.
      HT(JJ)=0.
      HN(JJ)=0.
      HE(JJ)=0.
      RR(JJ)=0.
      RC(JJ)=0.
      RT(JJ)=0.
      RN(JJ)=0.
      RE(JJ)=0.
   18 CONTINUE
      DO 1 JI=1,N
      DO 1 I=12,24,12
      Y=JI
      K=I
      DO 2 J=1,K
      X=RANF(Y)
    2 RAND(J)=X
      CALL GAUSS(K,JI,RORM)
      CALL CNORM(STD,PCT,K,AVG,RORM,CONT,JI)
      CALL TRIANG(A,K,TRIAN,JI)
      CALL EXPON(K,EXP,JI)
      CALL BOS(RAND,K,STATR)
```

```
      CALL BDS(CONT,K,STATC)
      CALL BDS(TRIAN,K,STATT)
      CALL BDS(RORM,K,STATN)
      CALL BDS(EXP,K,STATE)
      XR(K)=((STATR(1)-.5)**2)+XR(K)
      XT(K)=(STATT(1)**2)+XT(K)
      XC(K)=(STATC(1)**2)+XT(K)
      XN(K)=(STATN(1)**2)+XC(K)
      XE(K)=(STATE(1)**2)+XE(K)
      CALL SHIT(RAND,K,SHR)
      CALL SHIT(CONT,K,SHC)
      CALL SHIT(RORM,K,SHN)
      CALL SHIT(TRIAN,K,SHT)
      CALL SHIT(EXP,K,SHX)
      CALL ARANGE(RAND,K,RORD)
      CALL ARANGE(CONT,K,CORD)
      CALL ARANGE(RORM,K,RORD)
      CALL ARANGE(TRIAN,K,TORD)
      CALL ARANGE(EXP,K,EXPOD)
      CALL SMED(RORD,K,AMEDR)
      CALL SMED(CORD,K,AMEDC)
      CALL SMED(TORD,K,AMEDT)
      CALL SMED(RORD,K,AMEDN)
      CALL SMED(EXPOD,K,AMEDX)
      CALL HOGG(STATR,AMEDR,RORD,K,HCR)
      CALL HOGG(STATC,AMEDC,CORD,K,HCC)
      CALL HOGG(STATT,AMEDT,TORD,K,HCT)
      CALL HOGG(STATN,AMEDN,RORD,K,HCN)
      CALL HOGG(STATE,AMEDX,EXPOD,K,HCE)
      CALL AMID(RORD,K,RRG)
      CALL AMID(CORD,K,RCG)
      CALL AMID(TORD,K,RTG)
      CALL AMID(RORD,K,RNG)
      CALL AMID(EXPOD,K,REG)
      RR(K)=((RRG-.5)**2)+RR(K)
      RC(K)=(RCG**2)+RC(K)
      RT(K)=(RTG**2)+RT(K)
      RN(K)=(RNG**2)+RN(K)
      RE(K)=(REG**2)+RE(K)
      YR(K)=((AMEDR-.5)**2)+YR(K)
      YC(K)=(AMEDC**2)+YC(K)
      YT(K)=(AMEDT**2)+YT(K)
      YN(K)=(AMEDN**2)+YN(K)
      YE(K)=(AMEDX**2)+YE(K)
      HR(K)=((HCR-.5)**2)+HR(K)
      HC(K)=(HCC**2)+HC(K)
      HT(K)=(HCT**2)+HT(K)
      HN(K)=(HCN**2)+HN(K)
      HE(K)=(HCE**2)+HE(K)
      SR(K)=((SHR-.5)**2)+SR(K)
      SC(K)=(SHC**2)+SC(K)
      ST(K)=(SHT**2)+ST(K)
      SN(K)=(SHN**2)+SN(K)
      SE(K)=(SHX**2)+SE(K)
1     CONTINUE
```

```
      DO 20 KK=12,24,12
      RR(KK)=RR(KK)/N
      RC(KK)=RC(KK)/N
      RT(KK)=RT(KK)/N
      RN(KK)=RN(KK)/N
      RE(KK)=RE(KK)/N
      YR(KK)=YR(KK)/N
      YC(KK)=YC(KK)/N
      YT(KK)=YT(KK)/N
      YN(KK)=YN(KK)/N
      YE(KK)=YE(KK)/N
      XR(KK)=XR(KK)/N
      XC(KK)=XC(KK)/N
      XT(KK)=XT(KK)/N
      XN(KK)=XN(KK)/N
      XE(KK)=XE(KK)/N
      HR(KK)=HR(KK)/N
      HC(KK)=HC(KK)/N
      HT(KK)=HT(KK)/N
      HN(KK)=HN(KK)/N
      HE(KK)=HE(KK)/N
      SE(KK)=SE(KK)/N
      SN(KK)=SN(KK)/N
      ST(KK)=ST(KK)/N
      SC(KK)=SC(KK)/N
      SR(KK)=SR(KK)/N
      PRINT 502,N,YR(KK),YC(KK),YT(KK),YN(KK),YE(KK)
     $,HR(KK),HC(KK),HT(KK),HN(KK),HE(KK)
     $,XR(KK),XC(KK),XT(KK),XN(KK),XE(KK),RR(KK),RC(KK),RT(KK),RN(KK),
     $RE(KK),SR(KK),SC(KK),ST(KK),SN(KK),SE(KK)
20    CONTINUE
      CALL RANGET(Y)
      PUNCH 501,Y
501   FORMAT(F14.9)
500   FORMAT(F5.2,F5.2,F6.3,F5.2,I3)
502   FORMAT(28X*REPETITIONS=*I3,/,28X*VARIANCE ESTIMATE(MEAN SQUARE ERR
     $OR) FOR SAMPLE MEDIAN*//,28X,5F14.9,//,28X*FOR HODGS ESTIMATOR*//,
     $28X,5F14.9,//,28X*FOR SAMPLE MEAN*//,28X,5F14.9,//,28X*FOR MID RAN
     $GE*//,28X,5F14.9,//28X*FOR SWITZERS ESTIMATOR*//,28X,5F14.9)
510   FORMAT(1X,5F15.9)
600   FORMAT(1X,F15.9)
601   FORMAT(2X,4F15.9,I3)
      END
      SUBROUTINE AMID(X,K,B)
      DIMENSION X(46)
      B=(X(1)+X(K))/2.
      RETURN
      END

      SUBROUTINE SMED(ARRAY,K,AMED)
      DIMENSION ARRAY(43)
      N=K/2
      M=(K/2)+1
      AMED=(ARRAY(N)+ARRAY(M))/2.
      RETURN
      END
```

```
      SUBROUTINE ARANGE(X,K,RANGE)
      DIMENSION X(48),RANGE(48)
      NOP=K-1
      DO 10 I=1,NOP
      IP=I+1
      DO 10 J=IP,K
      IF(X(I).LE.X(J))GO TO 10
      TEMP=X(I)
      X(I)=X(J)
      X(J)=TEMP
   10 RANGE(I)=X(I)
      RANGE(K)=X(K)
      RETURN
      END

      SUBROUTINE SWIT(X,K,SW)
      DIMENSION A(8),B(8),C(8),X(48)
      L=K/6
      SQA=0.
      SQB=0.
      SQC=0.
      DO 50 I=1,L
      N=I*6
      M=N-5
      NN=N-1
      DO 56 II=M,NN
      JK=II+1
      DO 56 KK=JK,N
      IF(X(II).LE.X(KK))GO TO 56
      TEMP=X(II)
      X(II)=X(KK)
      X(KK)=TEMP
   56 X(II)=X(II)
      MN=M+1
      MM=N-1
      MI=M+2
      MJ=N-2
      A(I)=(X(N)+X(M))/2.
      B(I)=(X(MM)+X(MN))/2.
      C(I)=(X(MJ)+X(MI))/2.
      SQA=A(I)+SQA
      SQB=B(I)+SQB
   50 SQC=C(I)+SQC
      SQA=SQA/L
      SQB=SQB/L
      SQC=SQC/L
      DENOM=L*(L-1)
      SSQA=0.
      SSQB=0.
      SSQC=0.
      DO 51 JJ=1,L
        SSQA=(A(JJ)-SQA)**2+(SSQA)
        SSQB=(B(JJ)-SQB)**2+(SSQB)
   51 SSQC=(C(JJ)-SQC)**2+(SSQC)
      SSQA=SSQA/DENOM
      SSQB=SSQB/DENOM
      SSQC=SSQC/DENOM
      IF(SSQA.LE.SSQB)GO TO 52
      IF(SSQB.LE.SSQC)GO TO 54
   52 IF(SSQA.LE.SSQC)GO TO 53
      SW=SQC
      GO TO 55
   53 SW=SQA
      GO TO 55
   54 SW=SQB
   55 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE HOSG(X,B,C,K,A)
      DIMENSION X(8),C(48)
      X(8)=X(8)+3.
      IF(X(8).GT.4.)GO TO 35
      IF(X(8).LT.2.)GO TO 36
      A=X(1)
      GO TO 40
35    IF(X(8).LE.5.5)GO TO 37
      A=B
      GO TO 40
36    A=0.
      L=K/4
      M=K-(L-1)
      DO 38 I=M,K
38    A=C(I)+A
      DO 39 J=1,L
39    A=C(J)+A
      A=A/(2*L)
      GO TO 40
37    A=0.
      L=K/4
      M=(K/4)+1
      MM=K
      DO 34 N=M,MM
34    A=C(N)+A
      A=A/(K/2)
40    CONTINUE
      RETURN
      END

      SUBROUTINE GAUSS(K,JI,RORM)
      DIMENSION RORM(48)
      ZZ=JI
      DO 61 I=1,K
      X=0.
      DO 60 J=1,12
60    X=RANF(ZZ)+X
      X=X-6.
61    RORM(I)=X
      RETURN
      END

      SUBROUTINE TRIANG(A,K,TRIAN,JI)
      DIMENSION TRIAN(48)
      YY=JI*2
      DO 7 I=1,K
      X=0.
      DO 8 J=1,2
8     X=RANF(YY)+X
      TRIAN(I)=A*(X-1.)
7     CONTINUE
      RETURN
      END
```

```
      SUBROUTINE CNORM(STD,PCT,K,AVG,RORM,CONT,JI)
      DIMENSION CONT(48),RORM(48)
      YY=JI*3
      DO 4 L=1,K
      X=RANF(YX)
      IF(X.GT.PCT)GO TO 5
      IF(X.LE.PCT)GO TO 6
    5 CONT(L)=RORM(L)
      GO TO 4
    6 CONT(L)=AVG+STD*RORM(L)
    4 CONTINUE
      RETURN
      END

      SUBROUTINE SORT(A,K)
      DIMENSION A(1)
      LOGICAL SWITCH
      IF(K.EQ.1) RETURN
      J1=1
      J2=K-1
    1 SWITCH=.FALSE.
      DO 2 J=J1,J2
      IF(A(J).LE.A(J+1))GO TO 2
      T=A(J+1)
      A(J+1)=A(J)
      A(J)=T
      J4=J
      IF(SWITCH)GO TO 2
      J3=J
      SWITCH=.TRUE.
    2 CONTINUE
      IF(.NOT.SWITCH) RETURN
      J1=MAX0(1,J3-1)
      J2=MAX0(1,J4-1)
      GO TO 1
      END

      SUBROUTINE EXPON(K,A,JI)
      DIMENSION A(48)
      YZ=JI*4
      DO 25 I=1,K
      X=RANF(YZ)
      IF(X.LE..5)GO TO 26
      IF(X.GT..5)GO TO 27
   26 A(I)=ALOG(X)
      GO TO 28
   27 Y=(3./2.)-X
      A(I)=-ALOG(Y)
   28 CONTINUE
   25 CONTINUE
      RETURN
      END
```

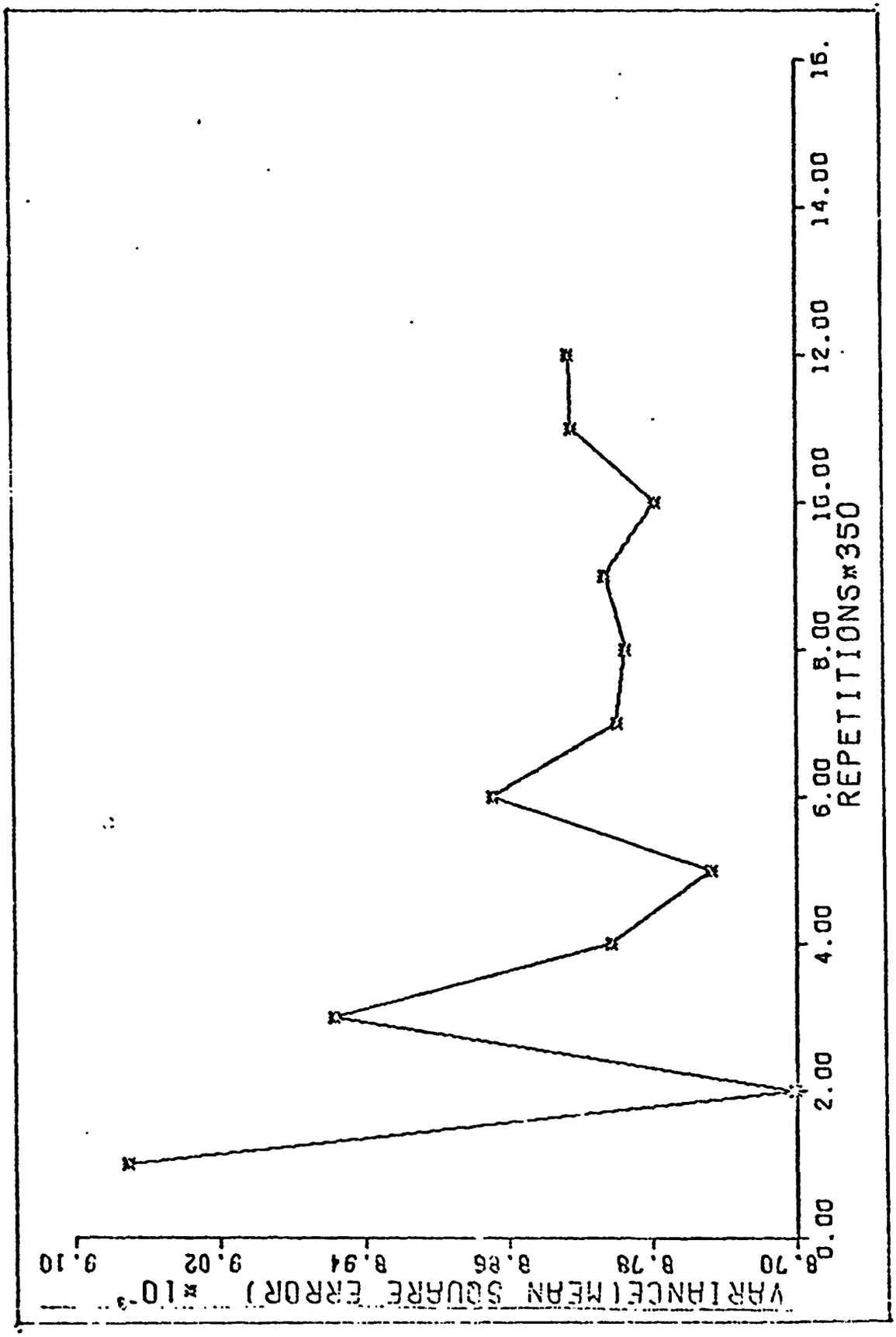
```
      SUBROUTINE AHOOG(A,K,ANS)
      DIMENSION A(48),B(48),C(12),ST(6,5),D(500)
      INTEGER ST
      DO 1 I=1,K
    1 B(I)=A(I)
      CALL SORT(B,K)
      I1=K/2
      DO 2 I=1,I1
    2 C(I)=B(I)+B(K-I+1)
      CALL SORT(C,K/2)
      C1=C(K/4)
      C2=C(K/4+1)
      IP=0
   12 N1=N2=N3=N4=N5=0
      DO 19 I=1,K
      DO 9 J=I,K
      T=B(I)+B(J)
      IF(T-C1)8,5,3
    3 IF(T-C2)7,6,4
    4 N5=N5+K-J+1
      GO TO 19
    5 N2=N2+1
      GO TO 9
    6 N4=N4+1
      GO TO 9
    7 N3=N3+1
      D(N3)=T
      GO TO 9
    8 N1=N1+1
    9 CONTINUE
   19 CONTINUE
      M=K*(K+1)/4
      IF(N3.EQ.0) GO TO 13
      IF(((N1+N2).LT.M).AND.((N4+N5).LT.M))GO TO 16
   24 IF((N1.LT.M).AND.(N5.LT.M))GO TO 18
   10 IP=IP+1
      C1=C(K/4-IP)
      C2=C(K/4+IP+1)
      GO TO 12
   13 IF((N1+N2).NE.(N4+N5))GO TO 24
   14 ANS=(C1+C2)*.25
      RETURN
   16 CALL SORT(D,N3)
      ID=M-N1-N2
      C1=D(ID)
      C2=D(ID+1)
      GO TO 14
   18 CALL SORT(D,N3)
      ID=M-N1-N2
      IF((N1+N2).GT.M)GO TO 21
      IF((N1+N2+N3).LT.M)GO TO 20
      IF(ID.NE.0) C1=D(ID)
      IF(ID.NE.N3) C2=D(ID+1)
      GO TO 14
   20 C1=C2

      GO TO 14
   21 C2=C1
      GO TO 14
      END
```

```
      SUBROUTINE GRAPH(Z,Y,N)
      DIMENSION Z(14),Y(14)
      CALL SCALE(Z,8.0,N,1)
      CALL SCALE(Y,5.0,N,1)
      CALL AXIS(0.0,0.0,15HREPETITIONS*350,-15,8.0,0.0,Z(N+1),Z(N+2))
      CALL AXIS(0.0,0.0,27HVARIANCE(MEAN SQUARE ERROR),27,5.0,90.0,Y(N+1
     S),Y(N+2))
      CALL LINE(Z,Y,N,1,1,11)
      CALL PLOT(10.,0.,-3)
      RETURN
      END
```

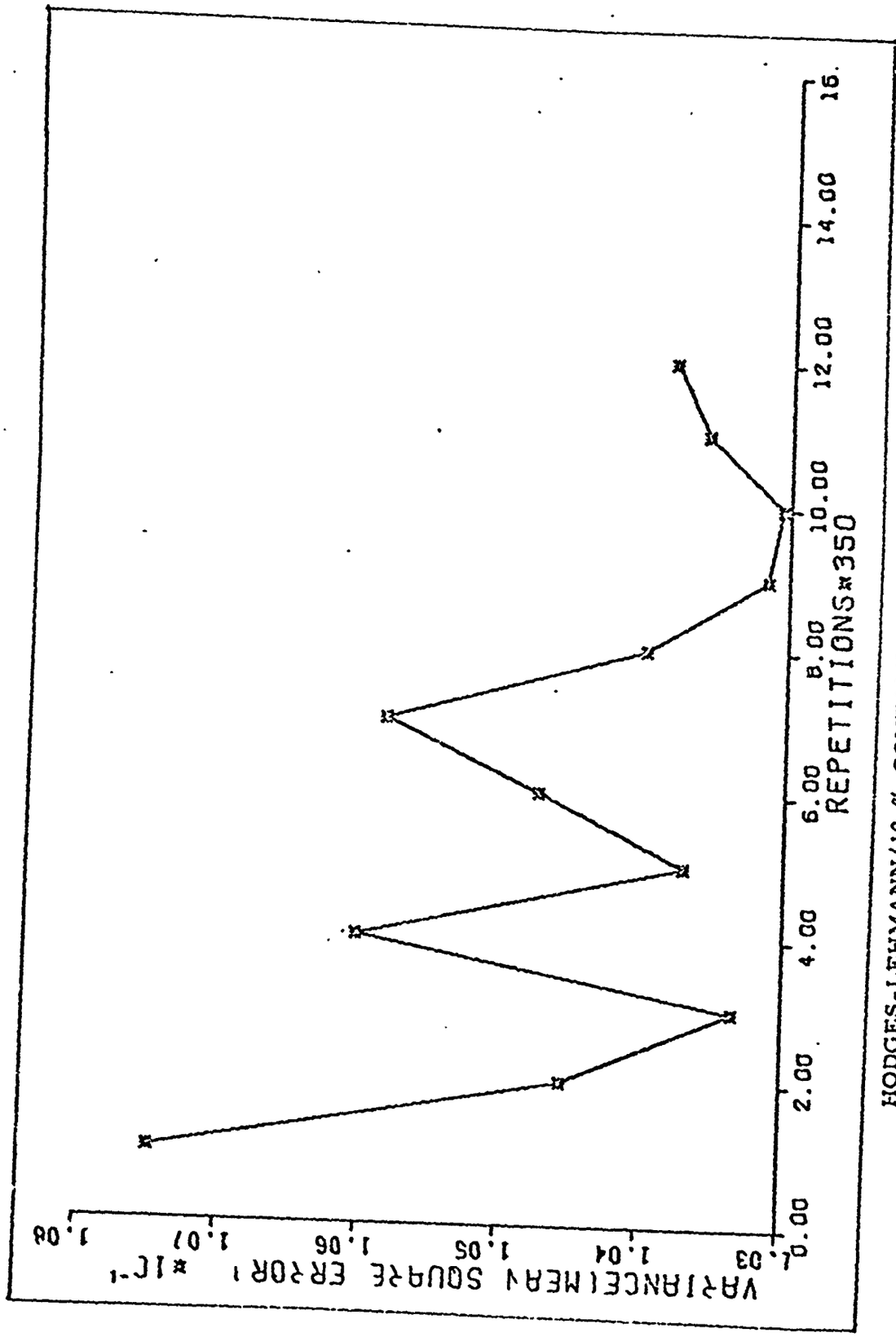
APPENDIX C

GRAPHS OF DATA

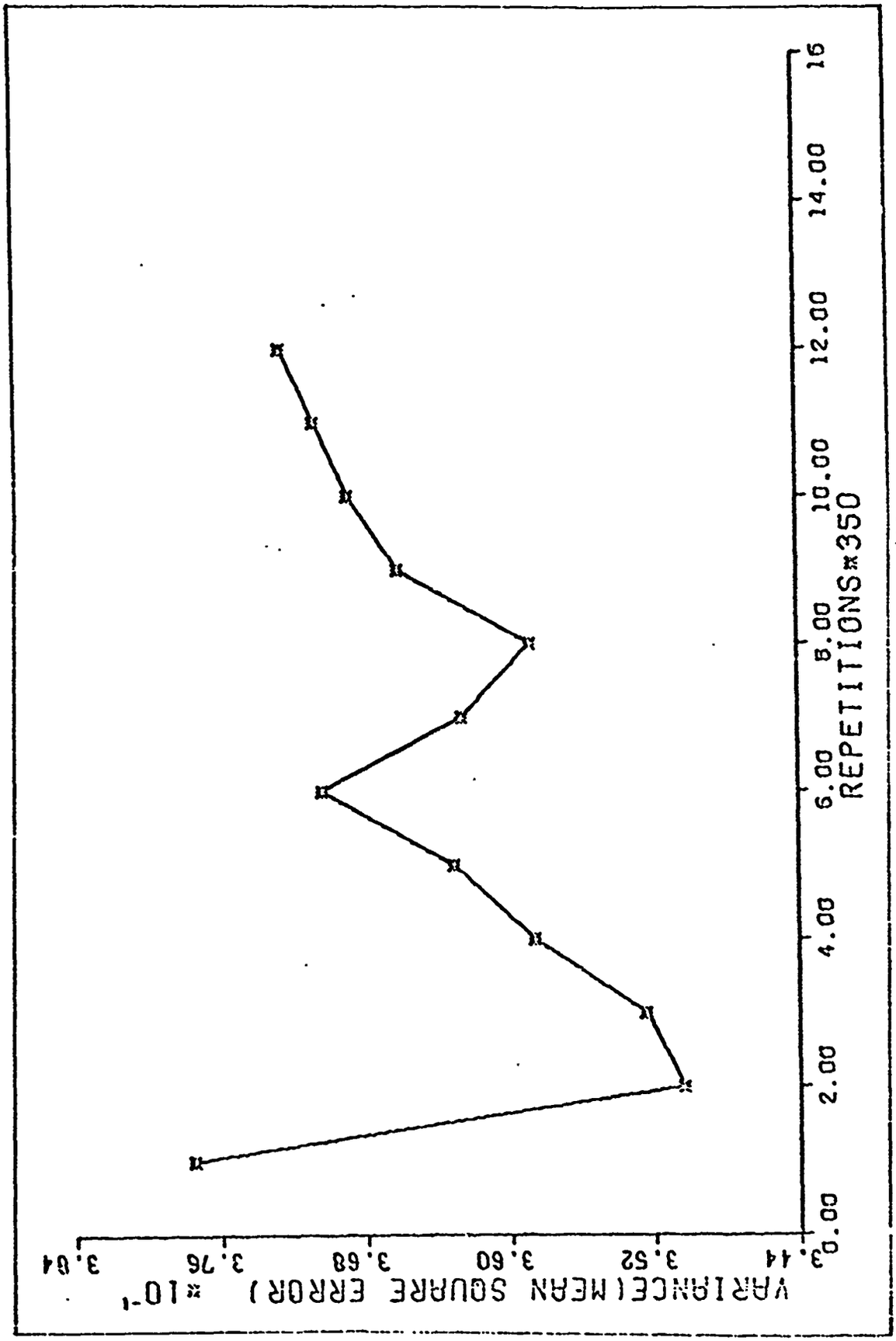
The thirty graphs presented in this section were selected from several hundred which were generated in the course of this investigation. One graph was chosen for each of the six estimators considered from each of the five distributions. The values plotted along the abscissa are the number of times the sample was drawn times 350. The values along the ordinate are the cumulative values for the mean square error. The graphs are labeled at the bottom by ESTIMATOR/PROBABILITY DISTRIBUTION/ SAMPLE SIZE.



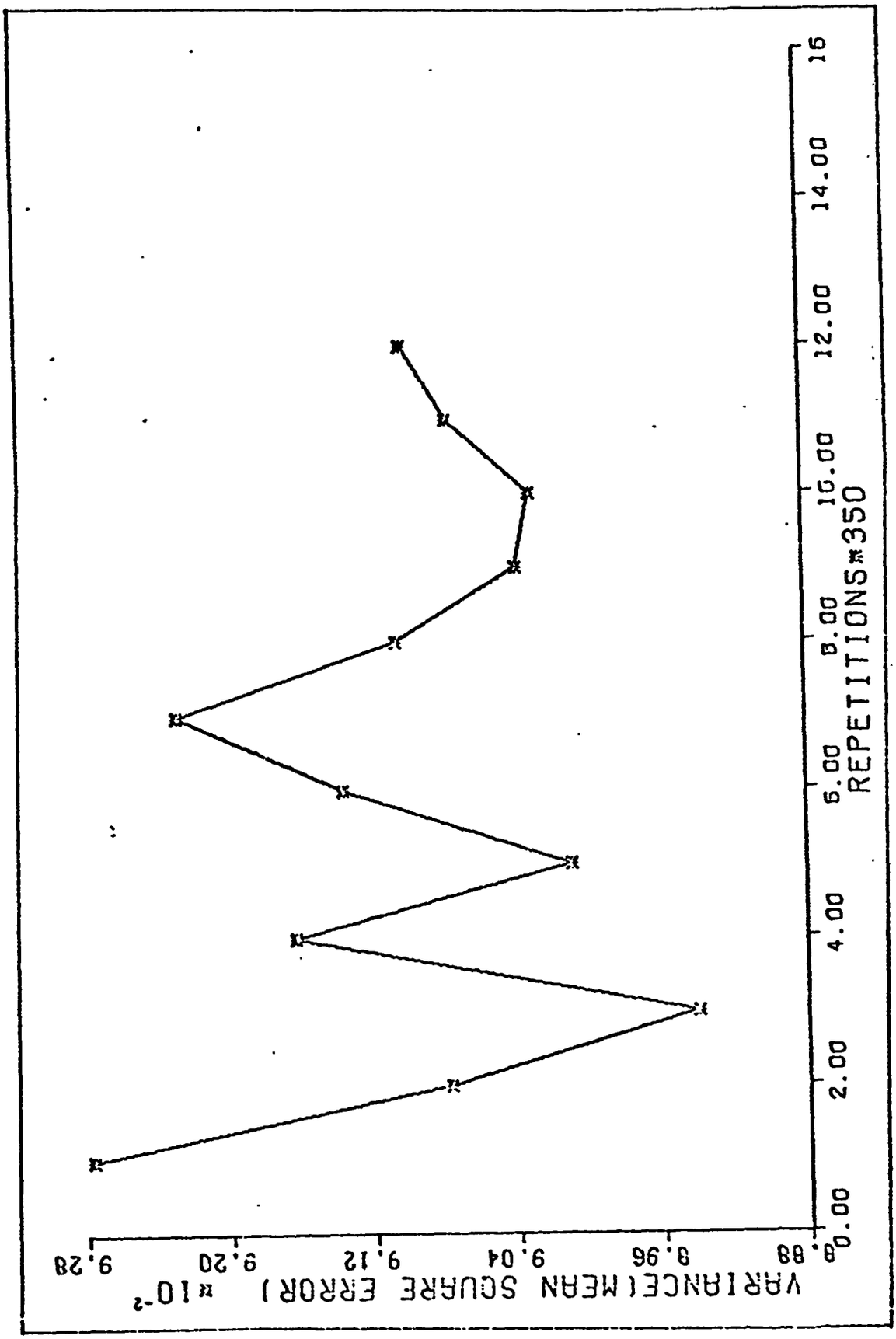
HODGES-LEHMANN/RECTANGULAR/12

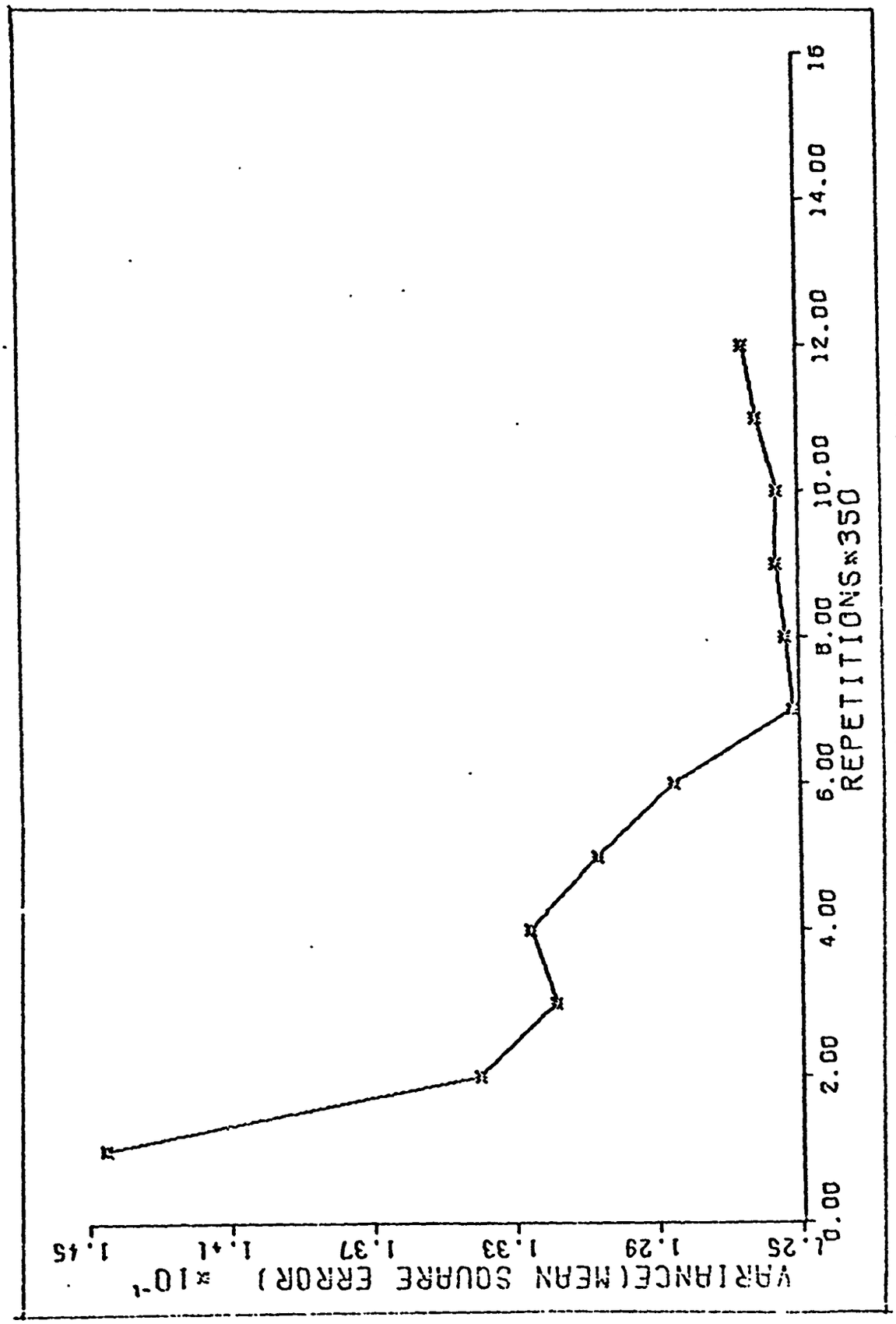


HODGES-LEHMANN/10 % CONTAMINATED NORMAL/12

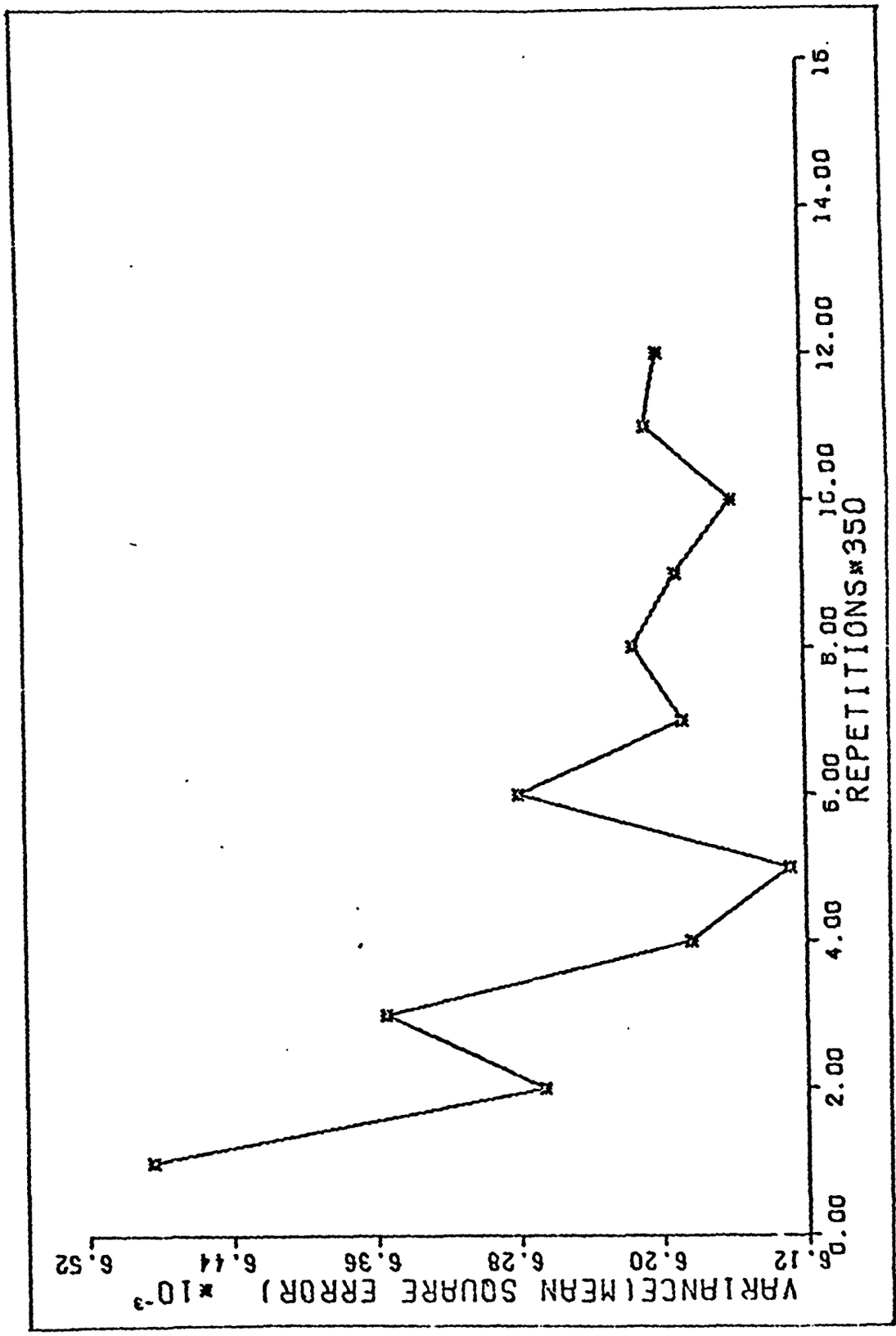


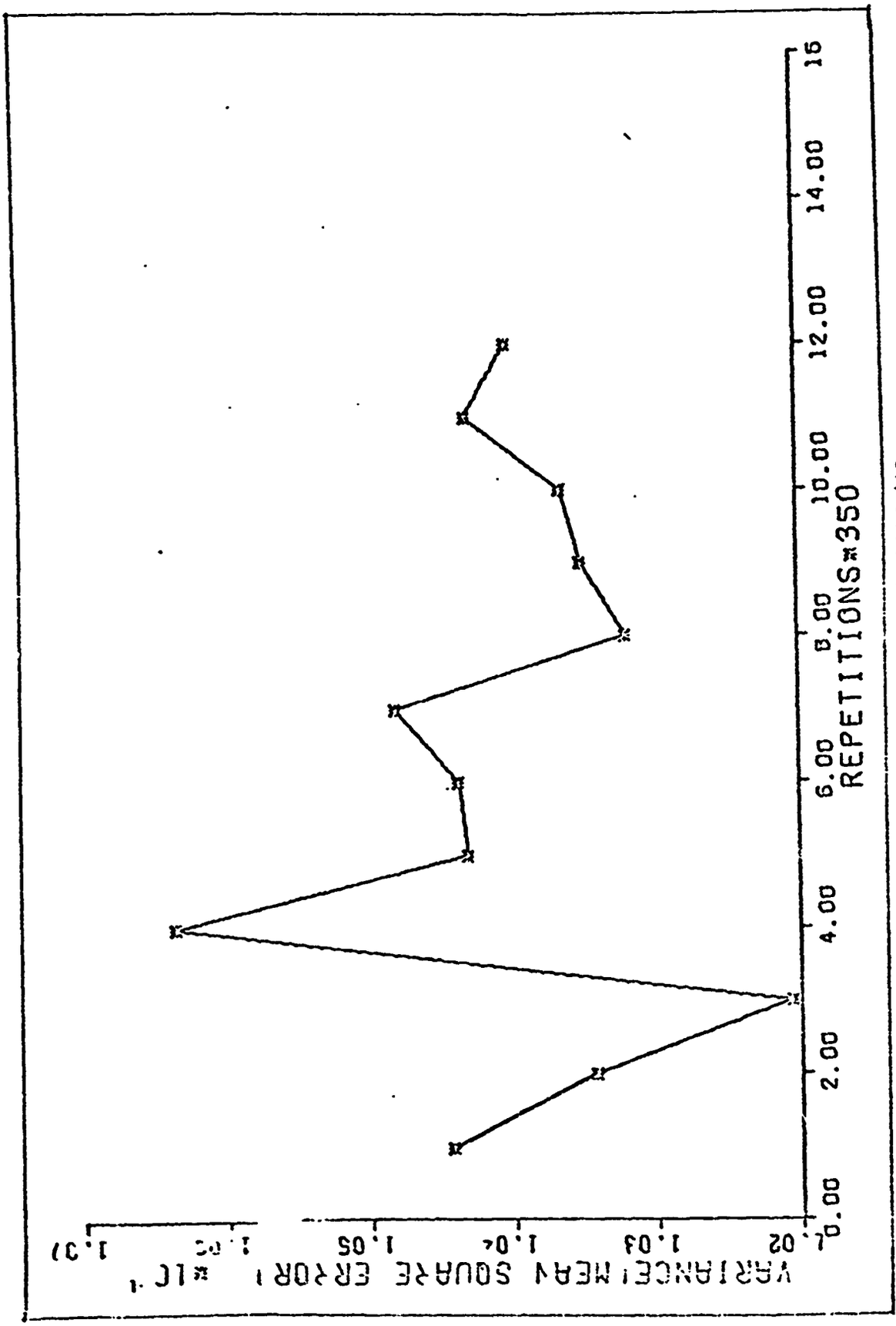
HODGES-LEHMANN/TRIANGULAR/12


VARIANCE(MEAN SQUARE ERROR) *10-2
8.88
8.96
9.04
9.12
9.20
9.28
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16
REPETITIONS=350
HODGES-LEHMANN/NORMAL(0,1)/12

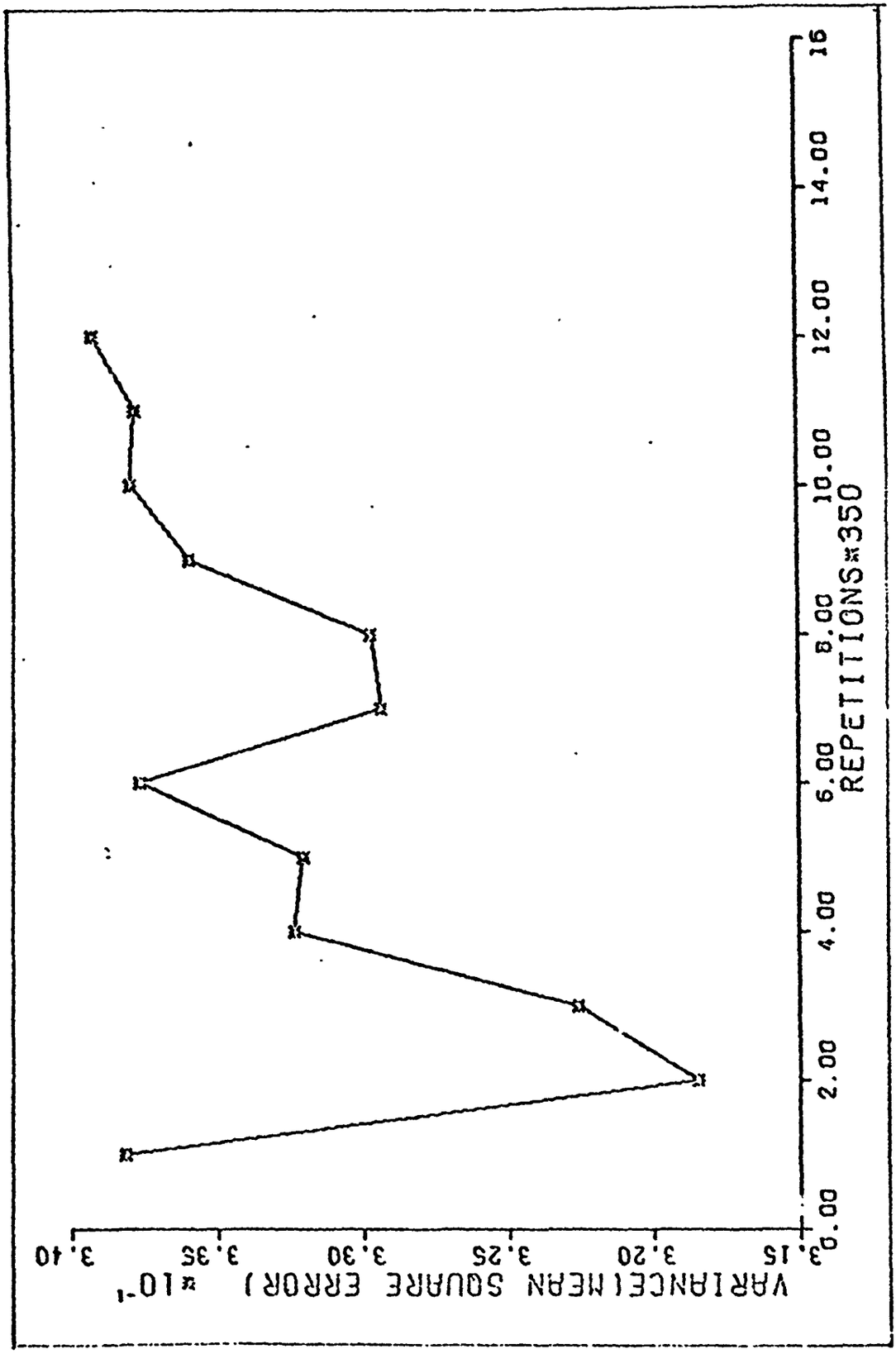


HODGES-LEHMANN/DOUB. EXPONENTIAL/12

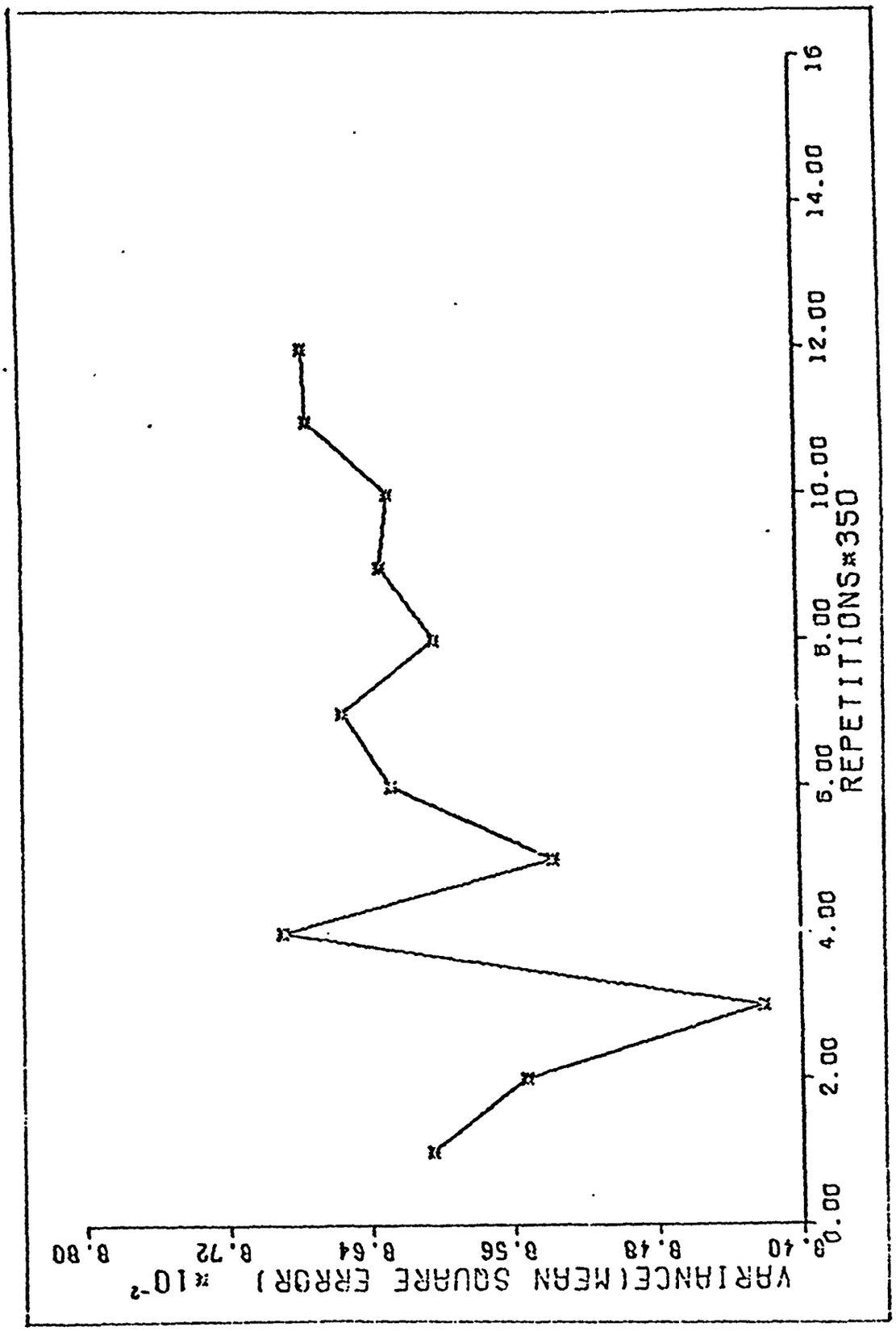

VARIANCE(MEAN SQUARE ERROR) *10^-3
6.12
6.20
6.28
6.36
6.44
6.52
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16.
REPETITIONS*350
HOGG'S/RECTANGULAR/12



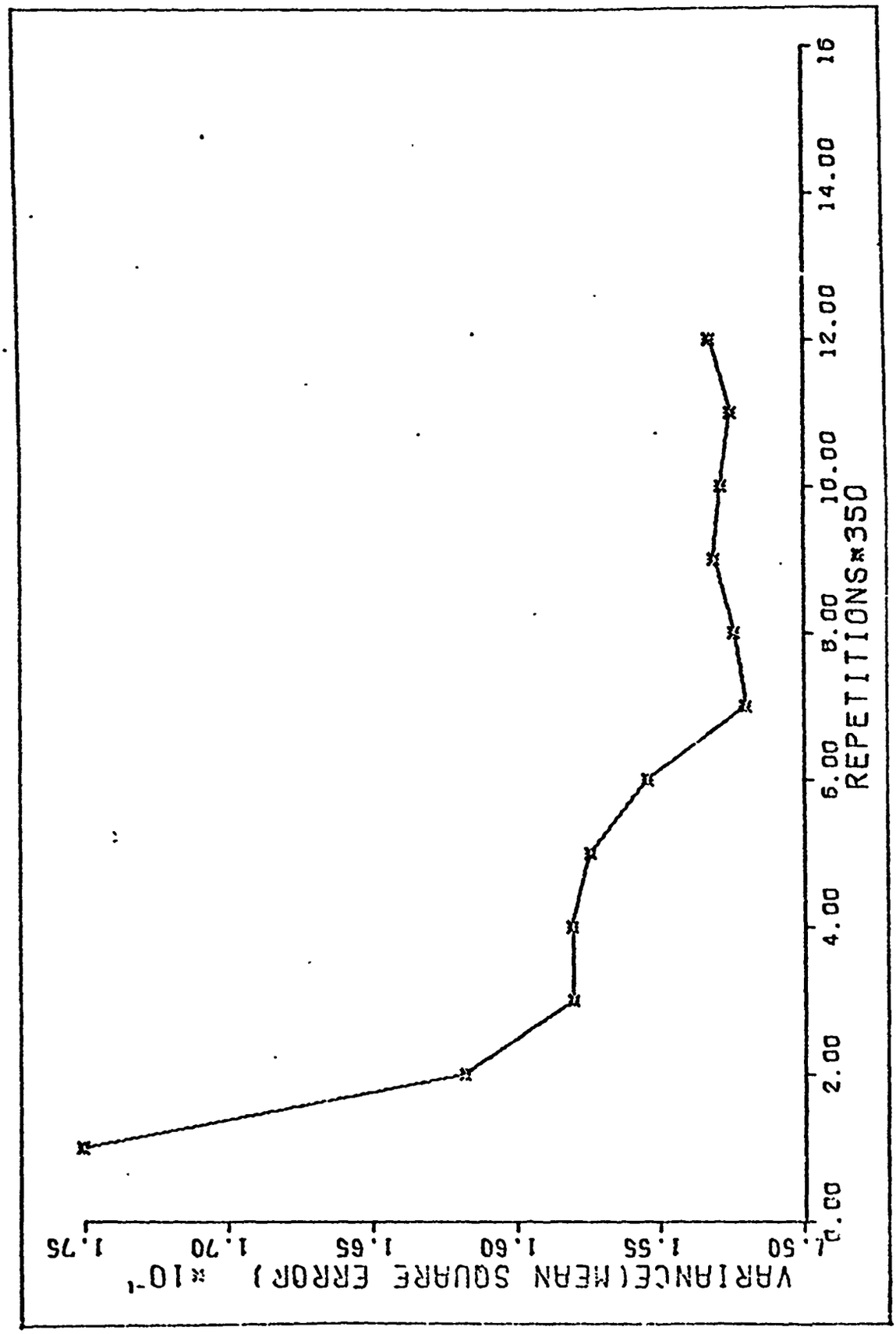
HOGG'S/10% CONTAMINATED NORMAL/12



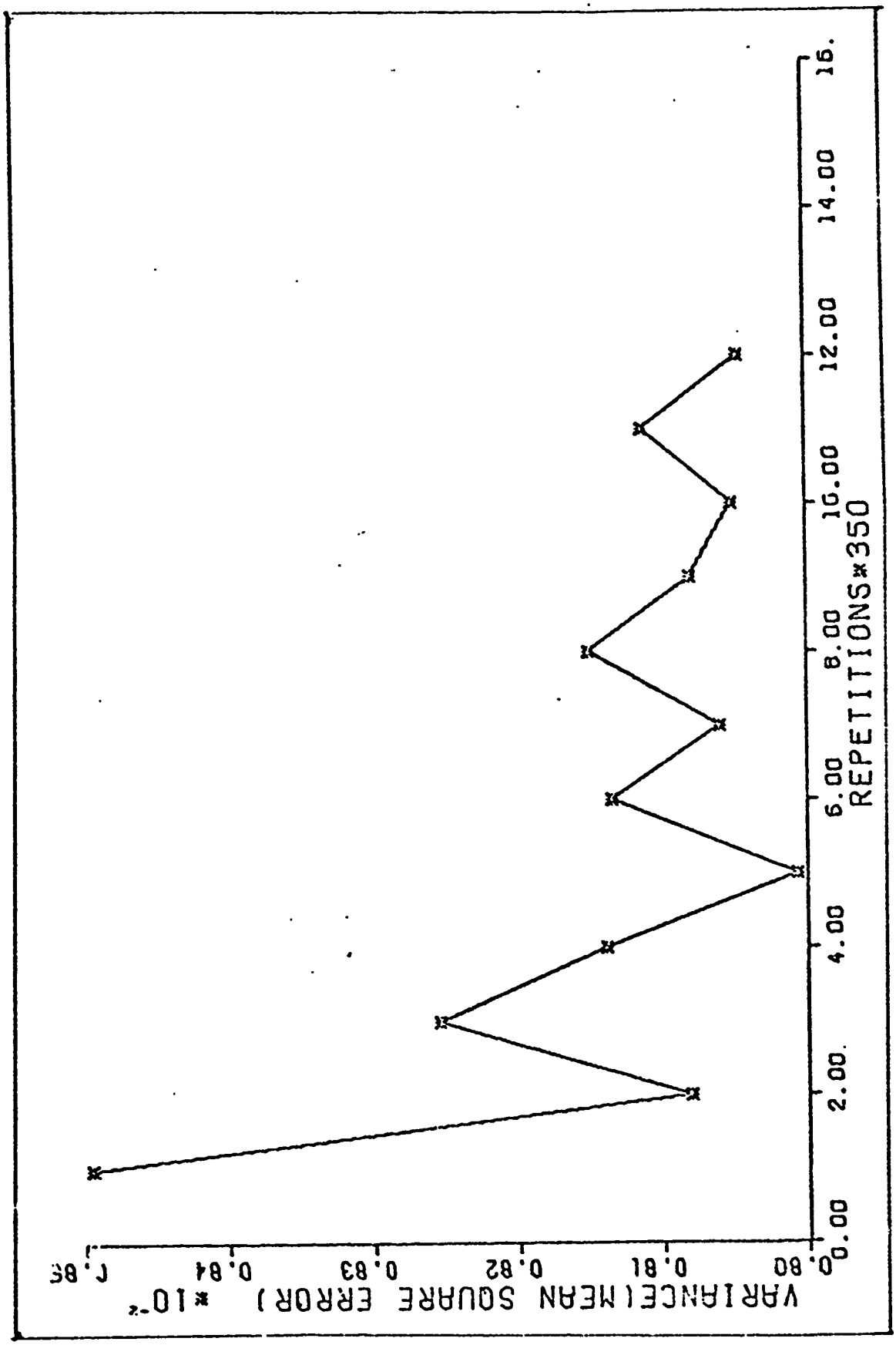
HOGG'S/TRIANGULAR/12

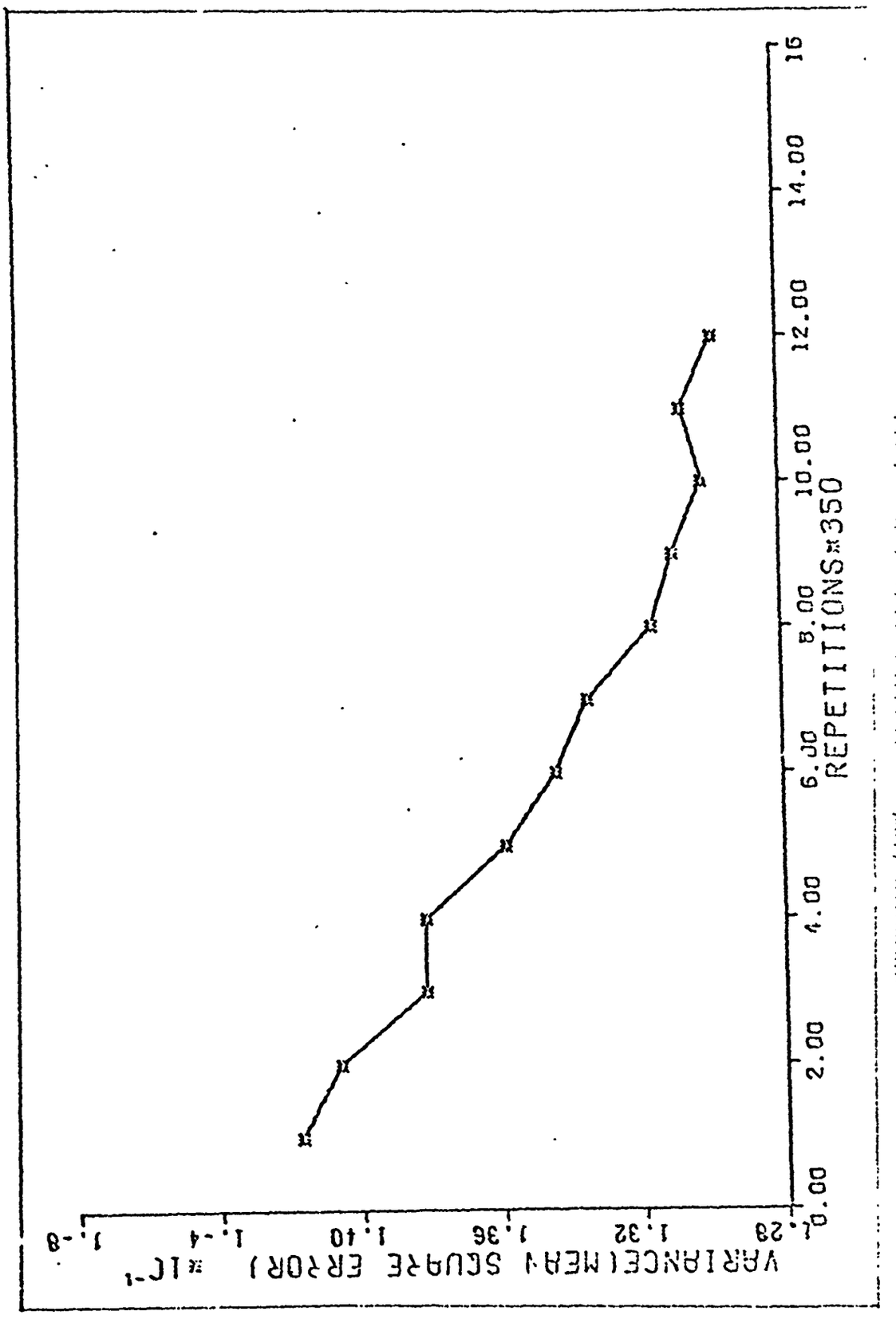


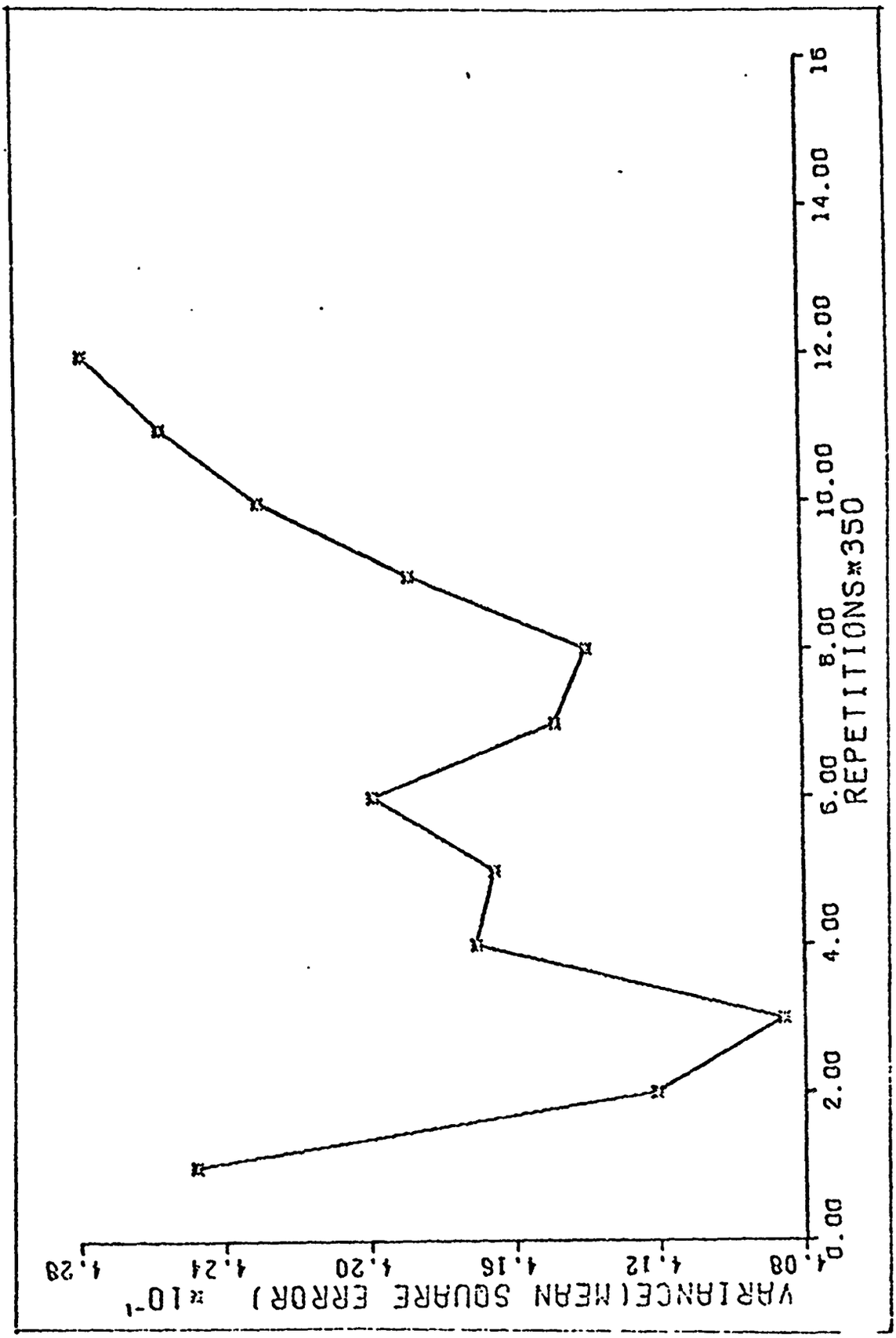
HOGG'S/NORMAL(0, 1)/12



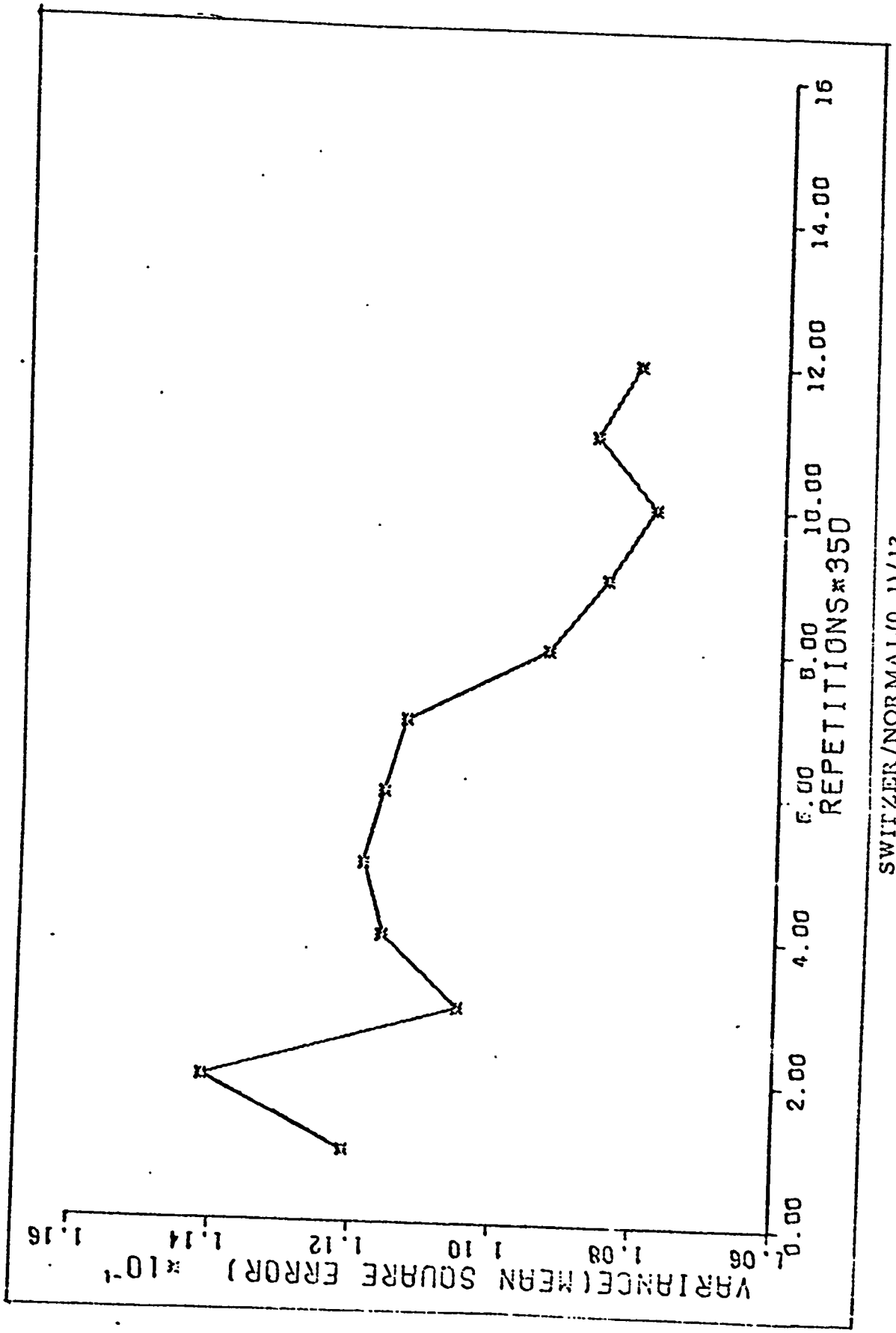
HOGG'S/DOUB. EXPONENTIAL/12

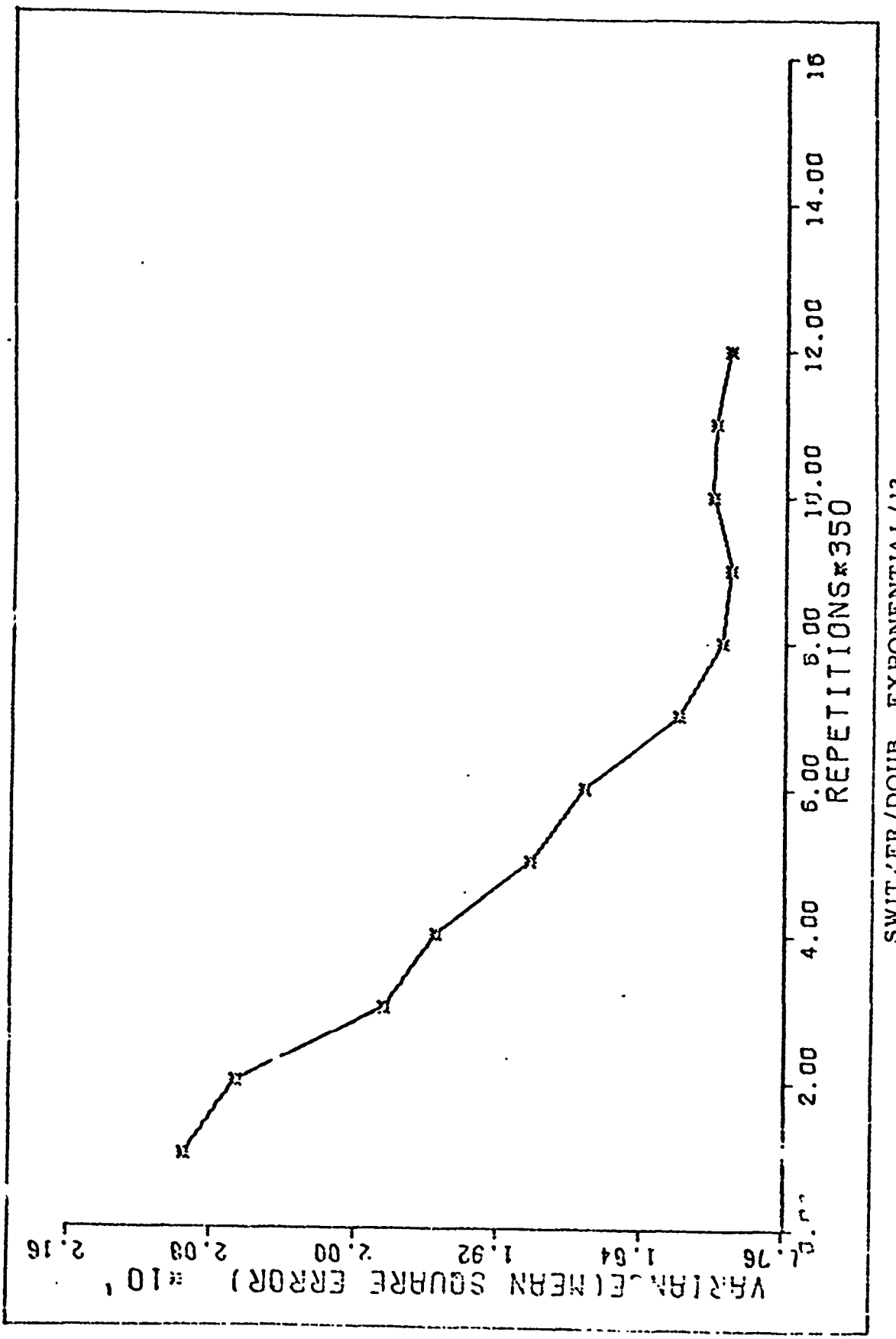

VARIANCE(MEAN SQUARE ERROR) *10⁻⁴
0.80
0.81
0.82
0.83
0.84
0.85
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16.
REPETITIONS*350
SWITZER/RECTANGULAR/12

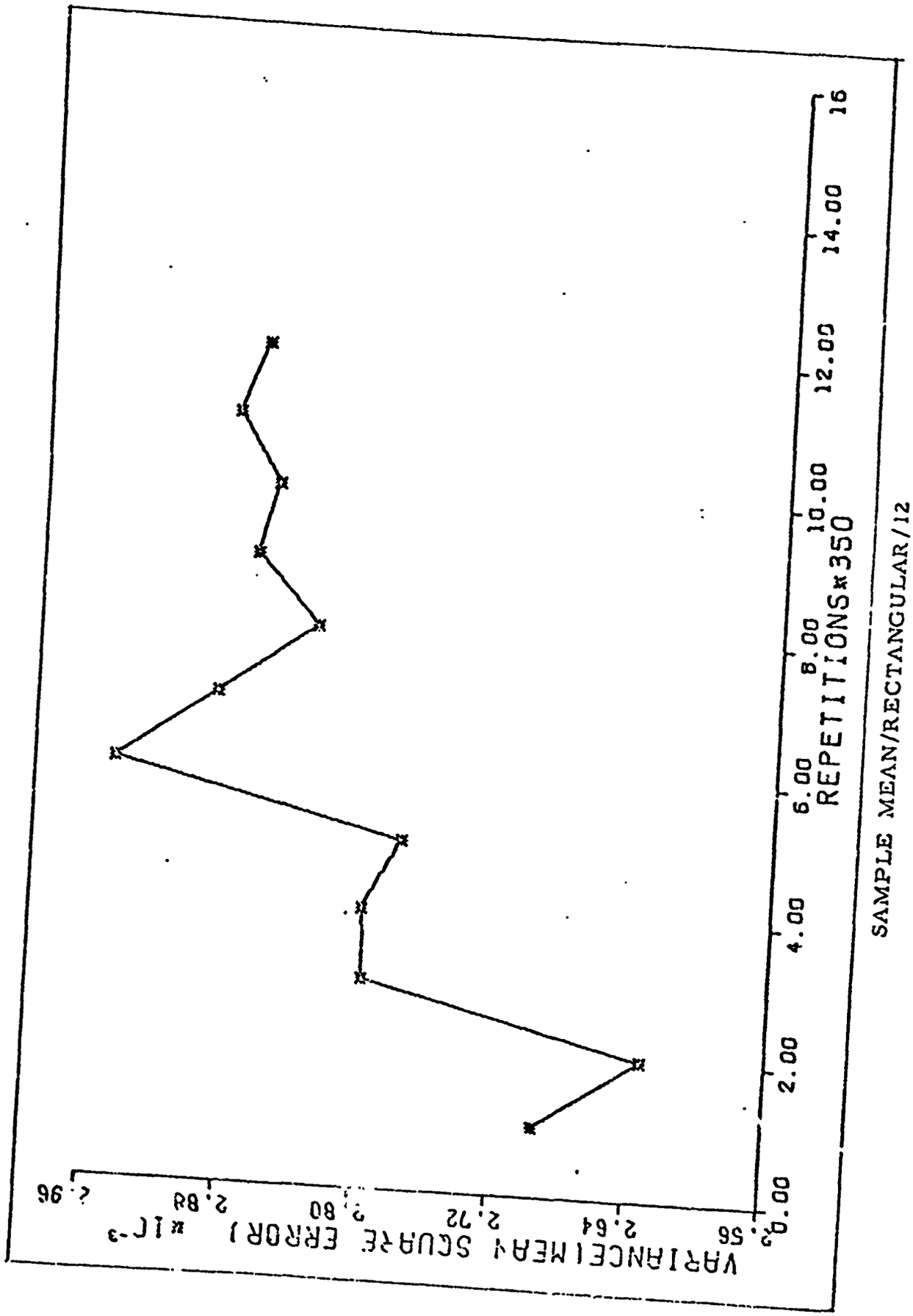

VARIANCE(MEAN SQUARE ERROR) *10^-1
1.28
1.32
1.36
1.40
1.-4
1.-8
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16
REPETITIONS*350



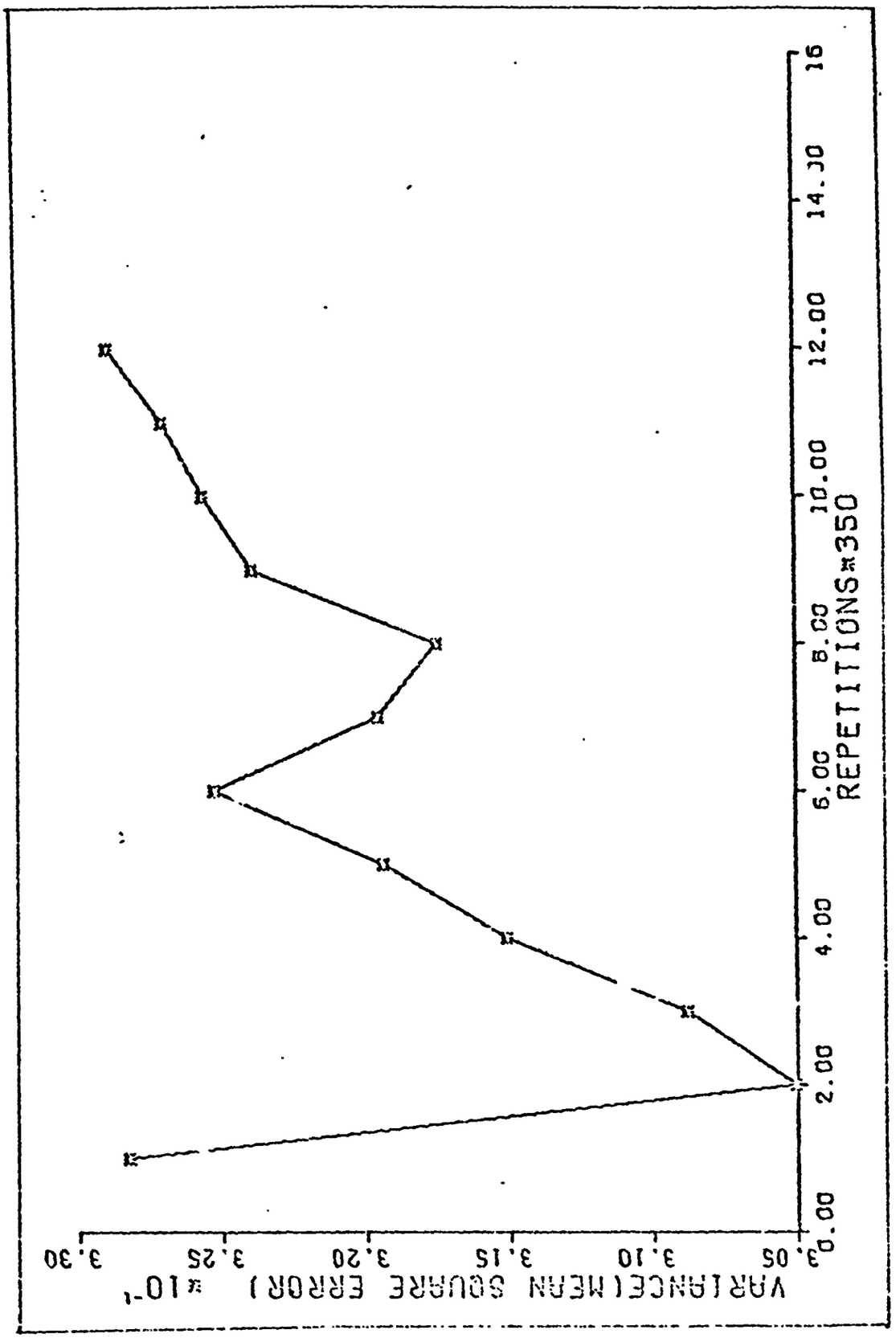
SWITZER/TRIANGULAR/12


VARIANCE(MEAN SQUARE ERROR) *10⁻¹
1.06
1.08
1.10
1.12
1.14
1.16
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16
REPETITIONS*350
SWITZER/NORMAL(0, 1)/12


VARIANCE(MEAN SQUARE ERROR) *10'
1.76
1.64
1.92
2.00
2.03
2.16
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16
REPETITIONS*350
SWITZER/DOUB. EXPONENTIAL/12

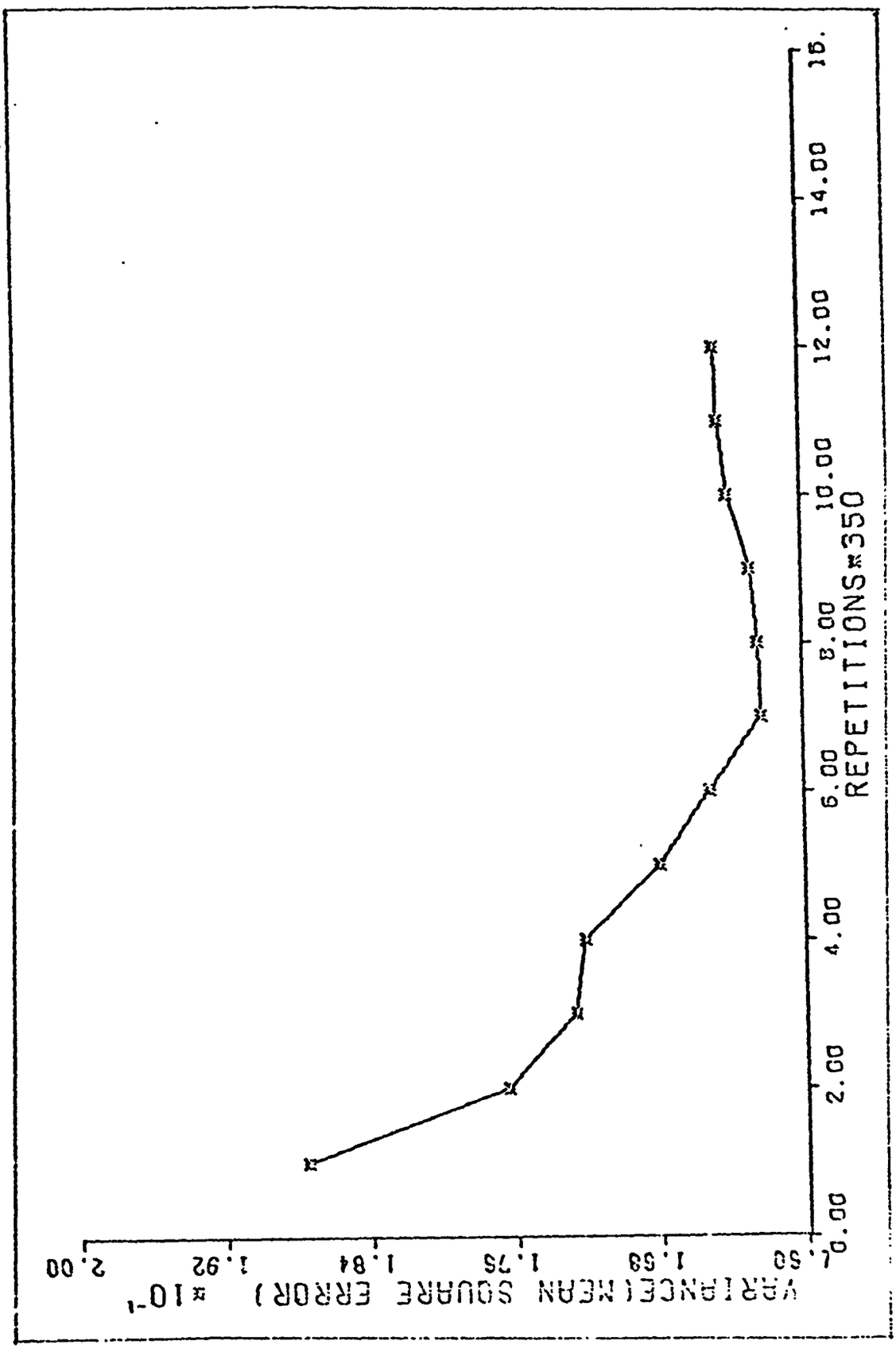

VARIANCE(MEAN SQUARE ERROR) *1E-3
2.56
2.64
2.72
2.80
2.88
2.96
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16
REPETITIONS*350
SAMPLE MEAN/RECTANGULAR/12

VARIANCE(MEAN SQU E ERROR) *10-1
1.00
1.02
1.04
1.06
1.08

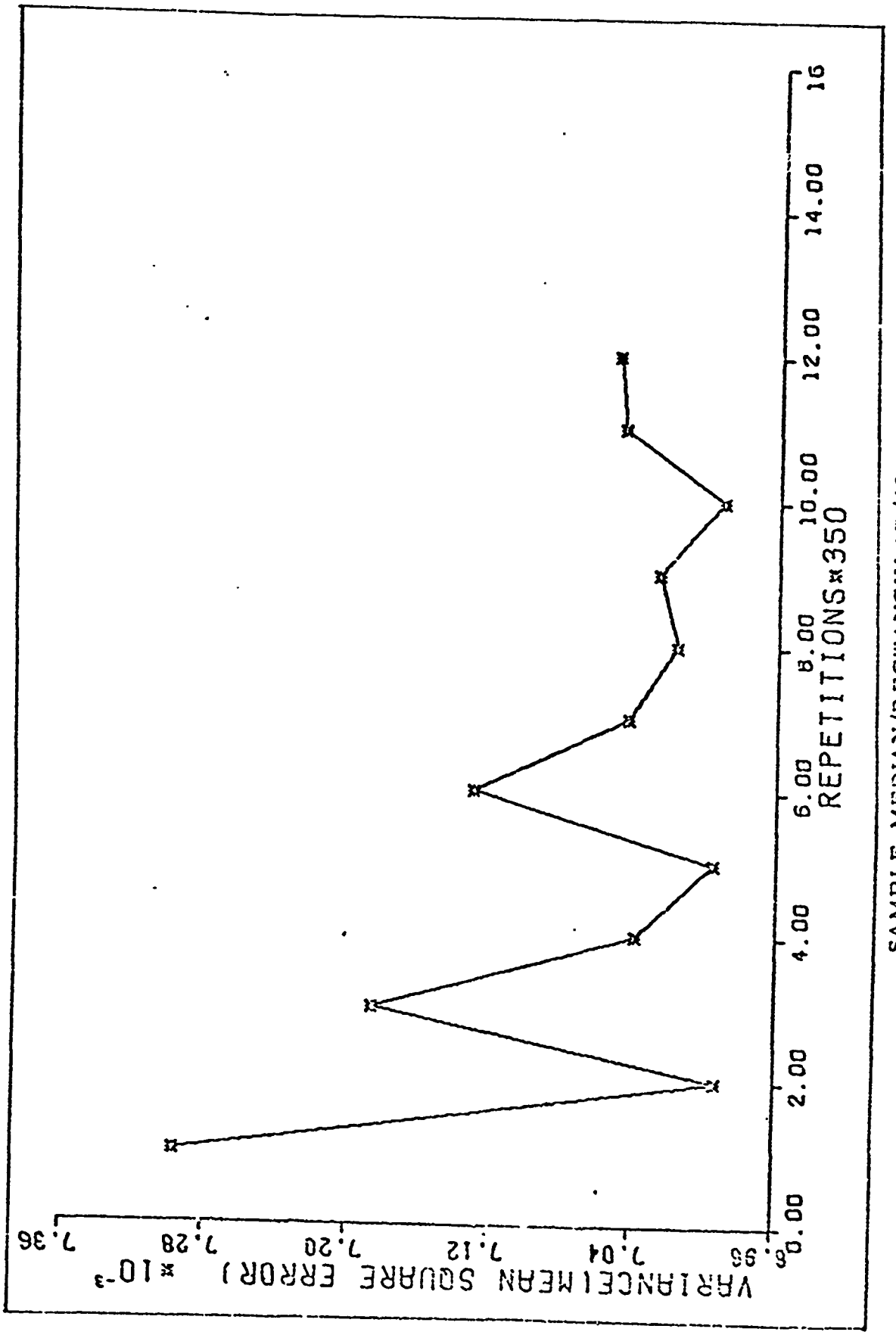


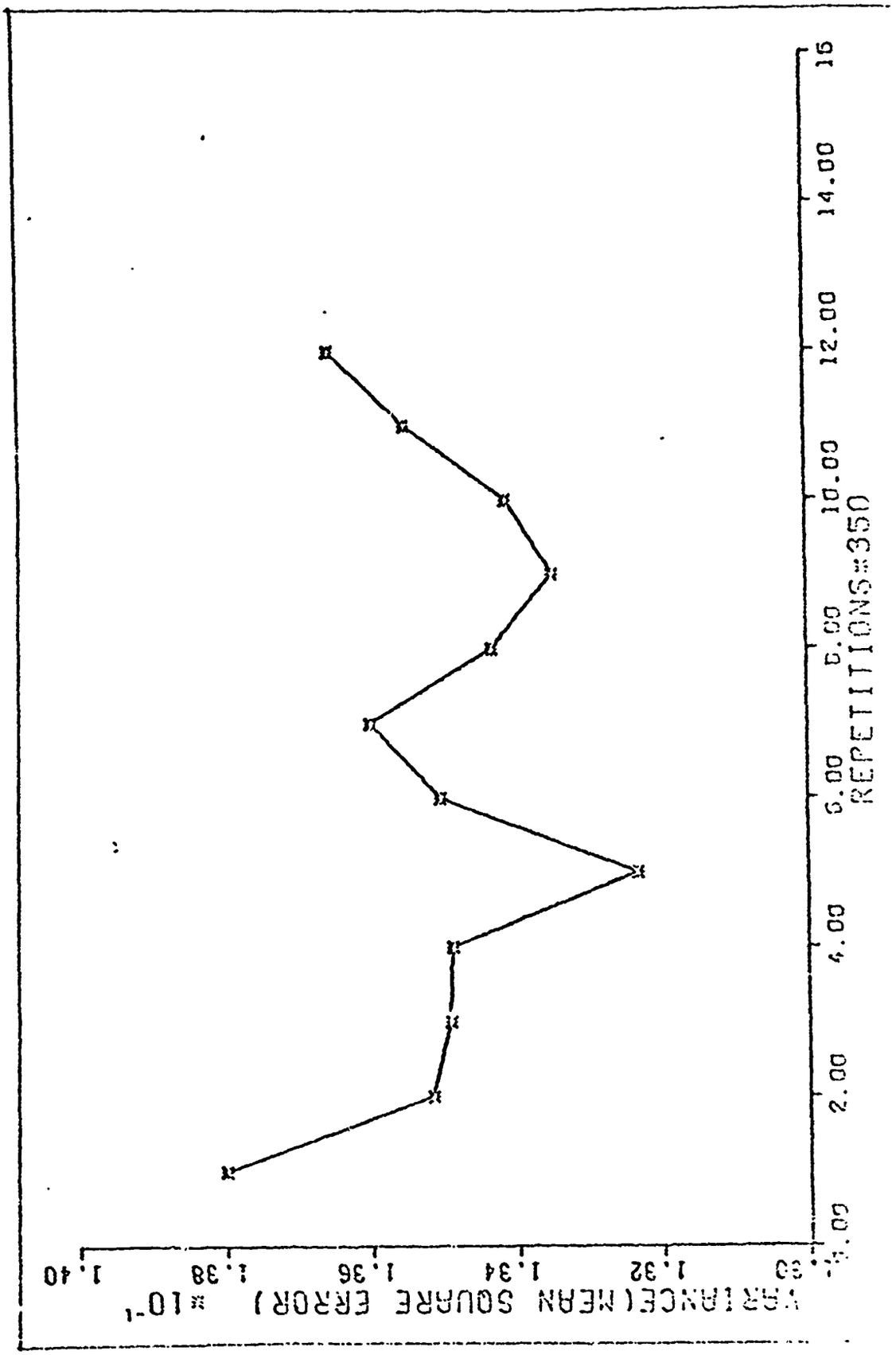
SAMPLE MEAN/TRIANGULAR/12

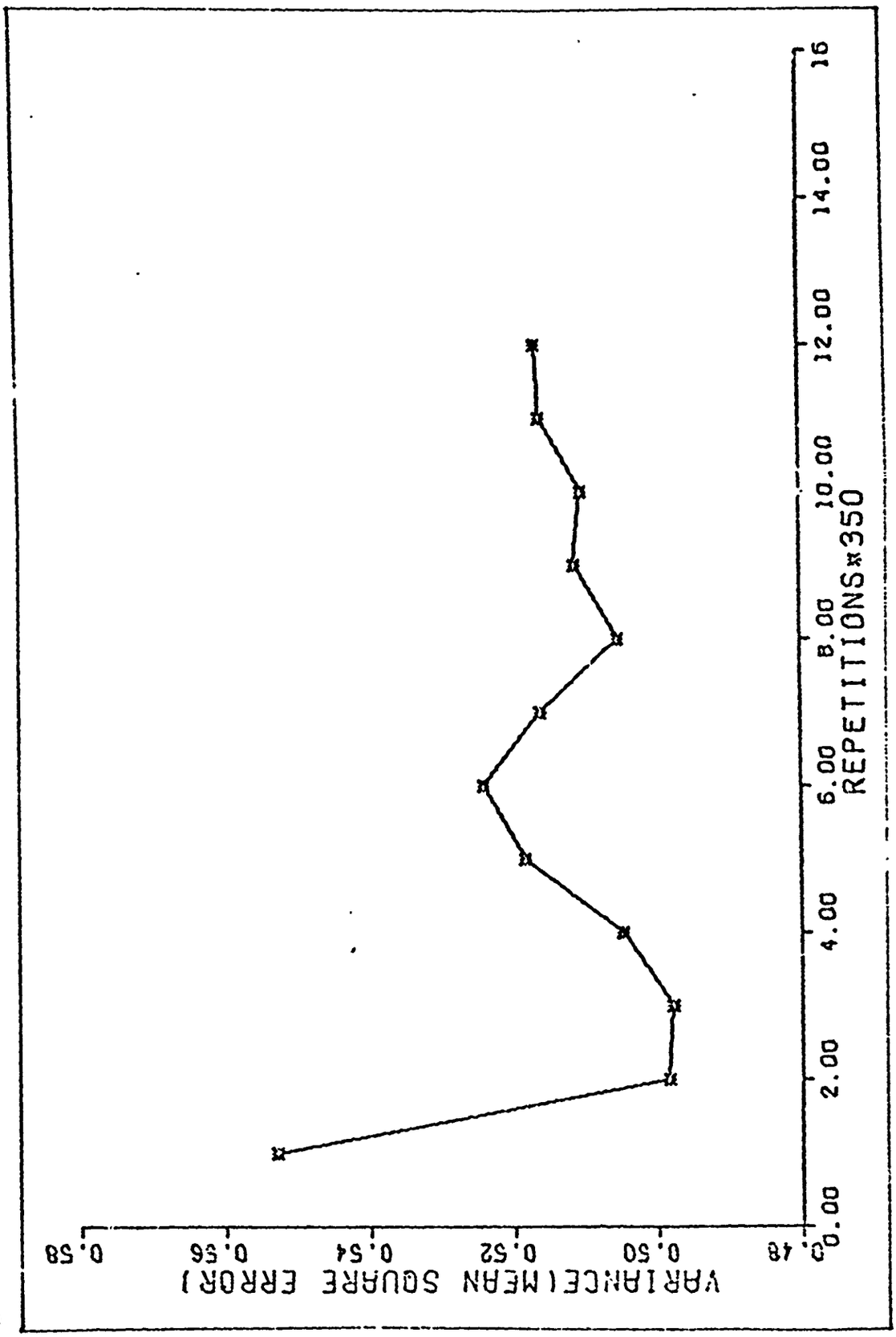
SAMPLE MEAN/NORMAL(0, 1)/12



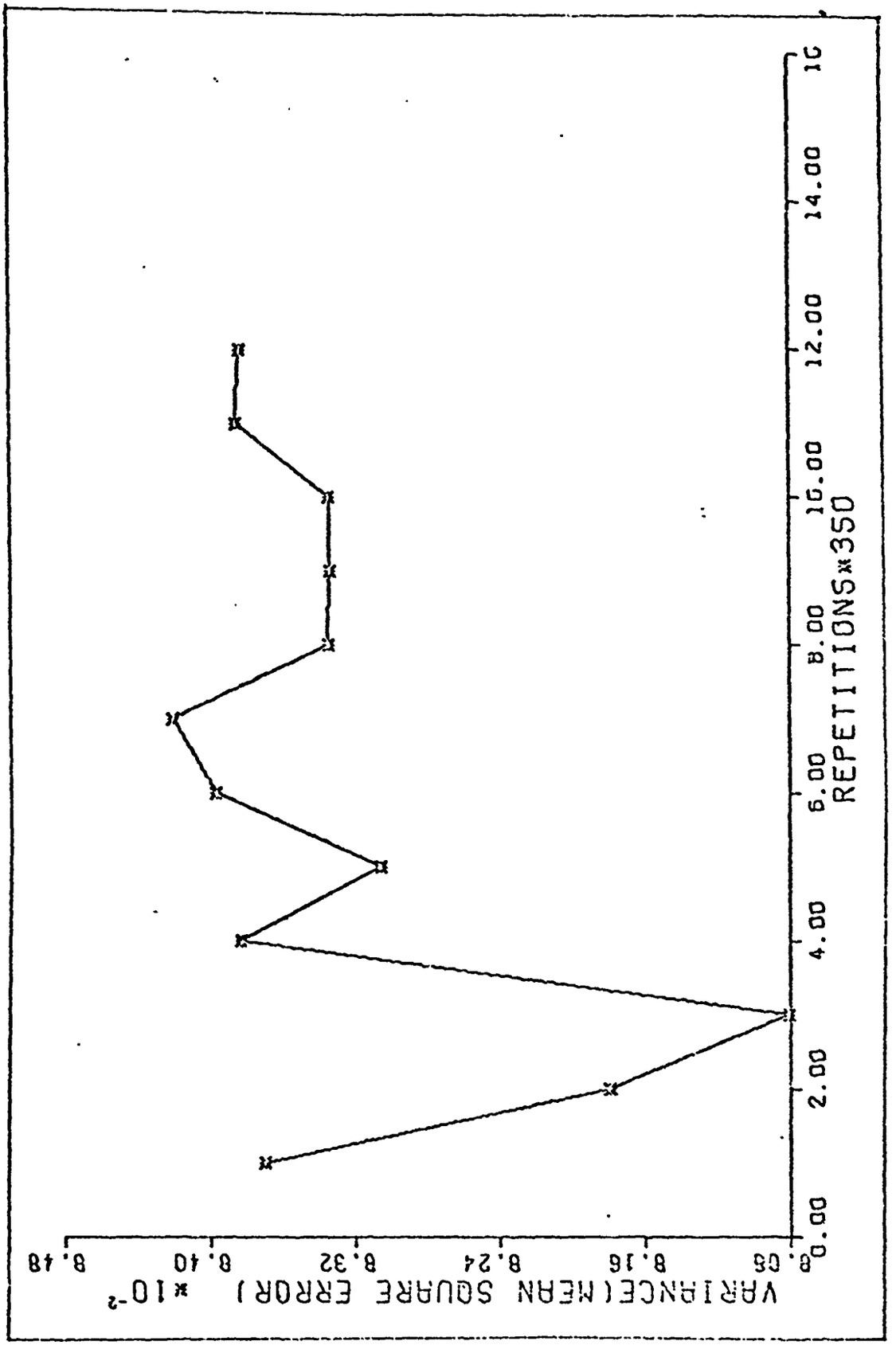
SAMPLE MEAN/DOUB. EXPONENTIAL/12


VARIANCE(MEAN SQUARE ERROR) *10-3
6.95
7.04
7.12
7.20
7.28
7.36
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16
REPETITIONS*350
SAMPLE MEDIAN/RECTANGULAR/12

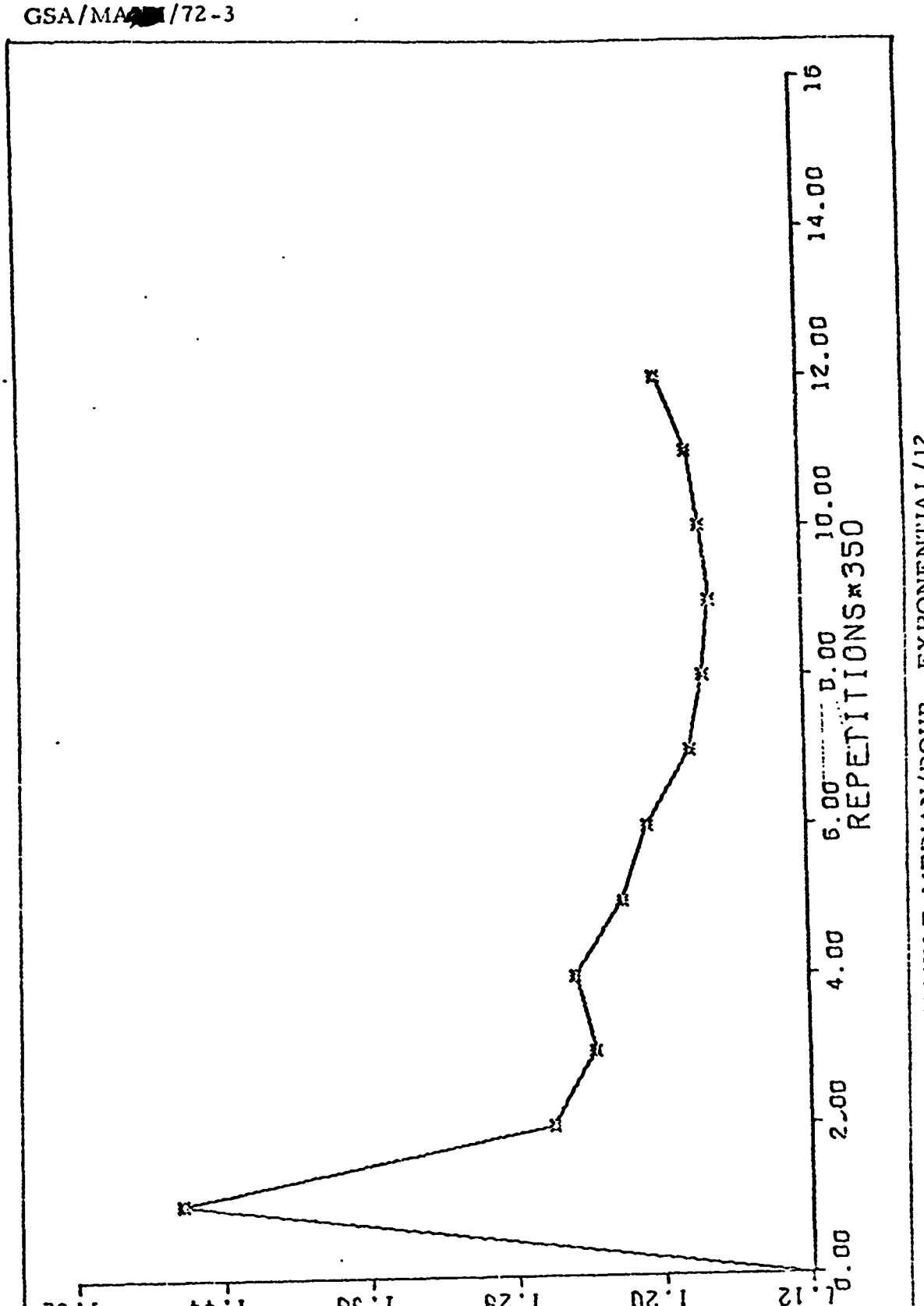

VARIANCE(MEAN SQUARE ERROR) *10⁻¹
1.30
1.32
1.34
1.36
1.38
1.40
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16
REPETITIONS=350



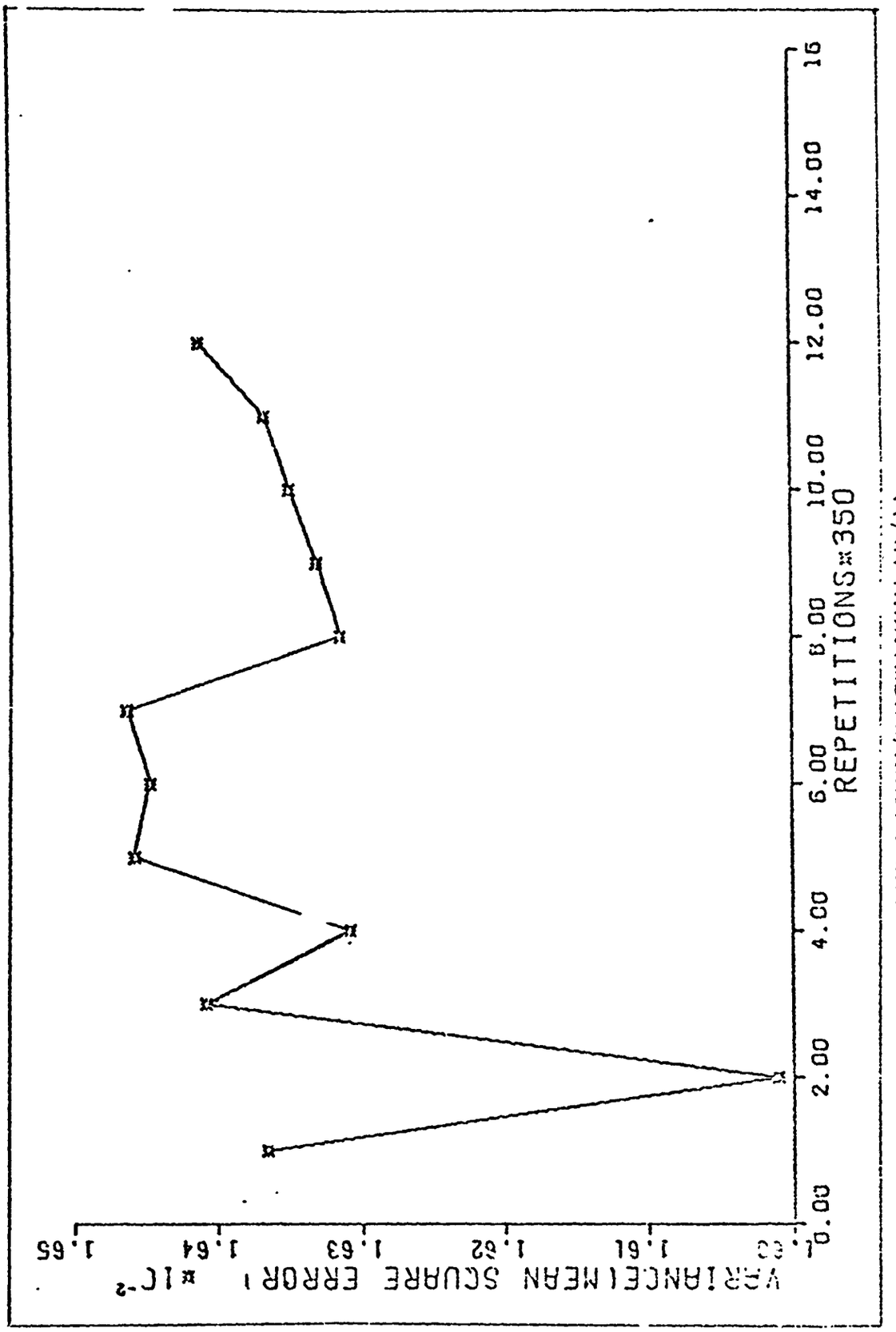
SAMPLE MEDIAN/TRIANGULAR/12



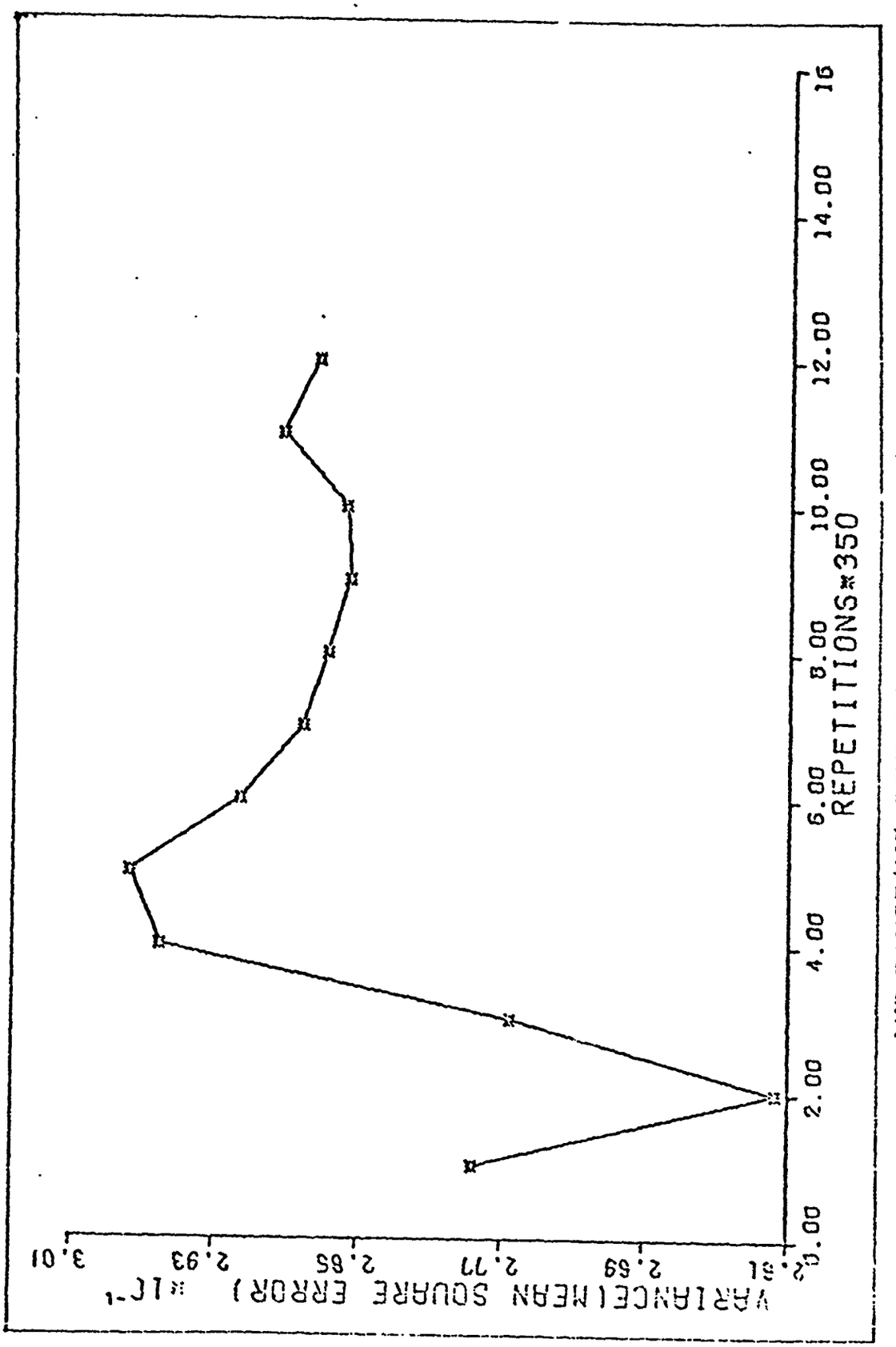
SAMPLE MEDIAN/NORMAL(0,1)/12



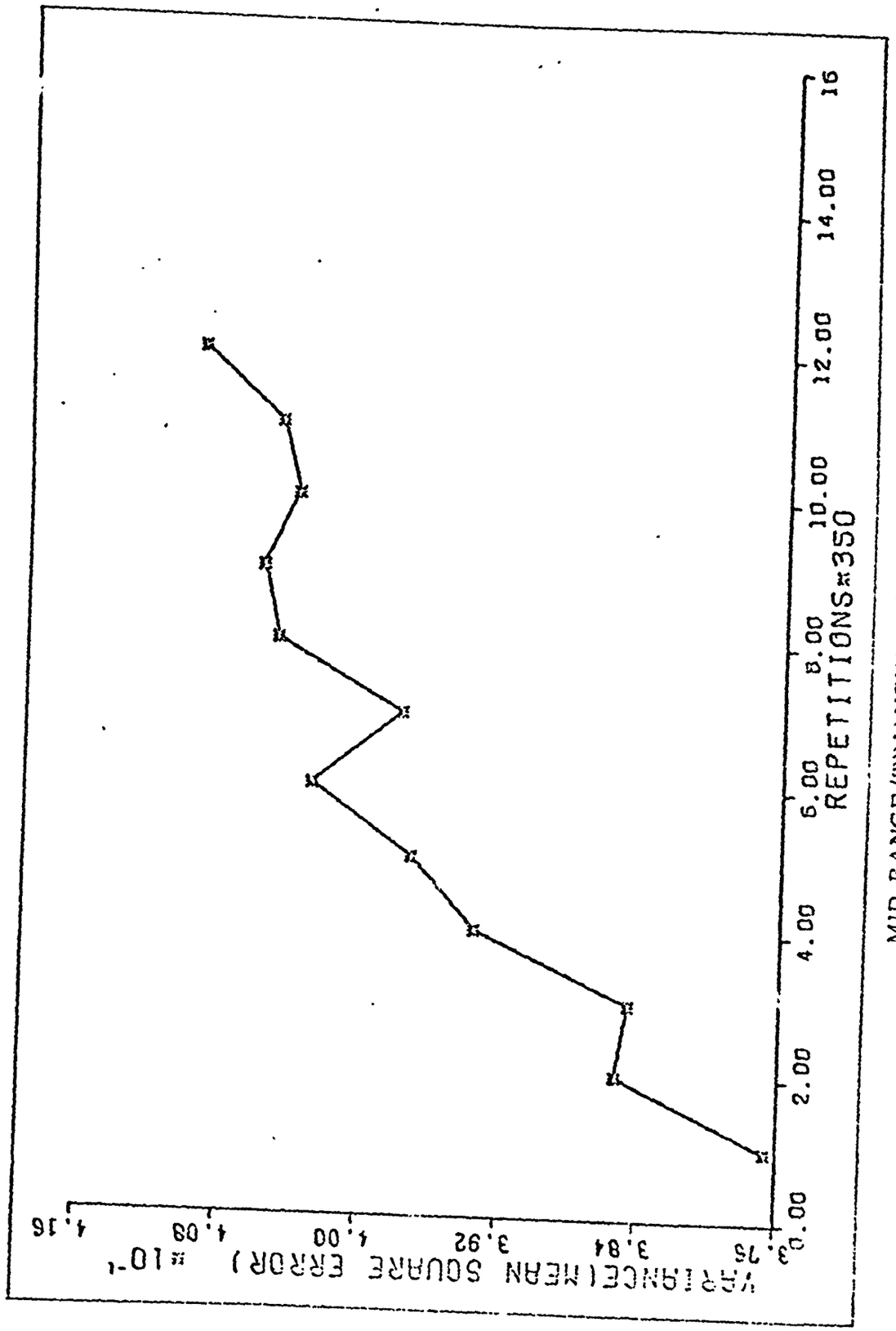
SAMPLE MEDIAN/DOUB. EXPONENTIAL/12



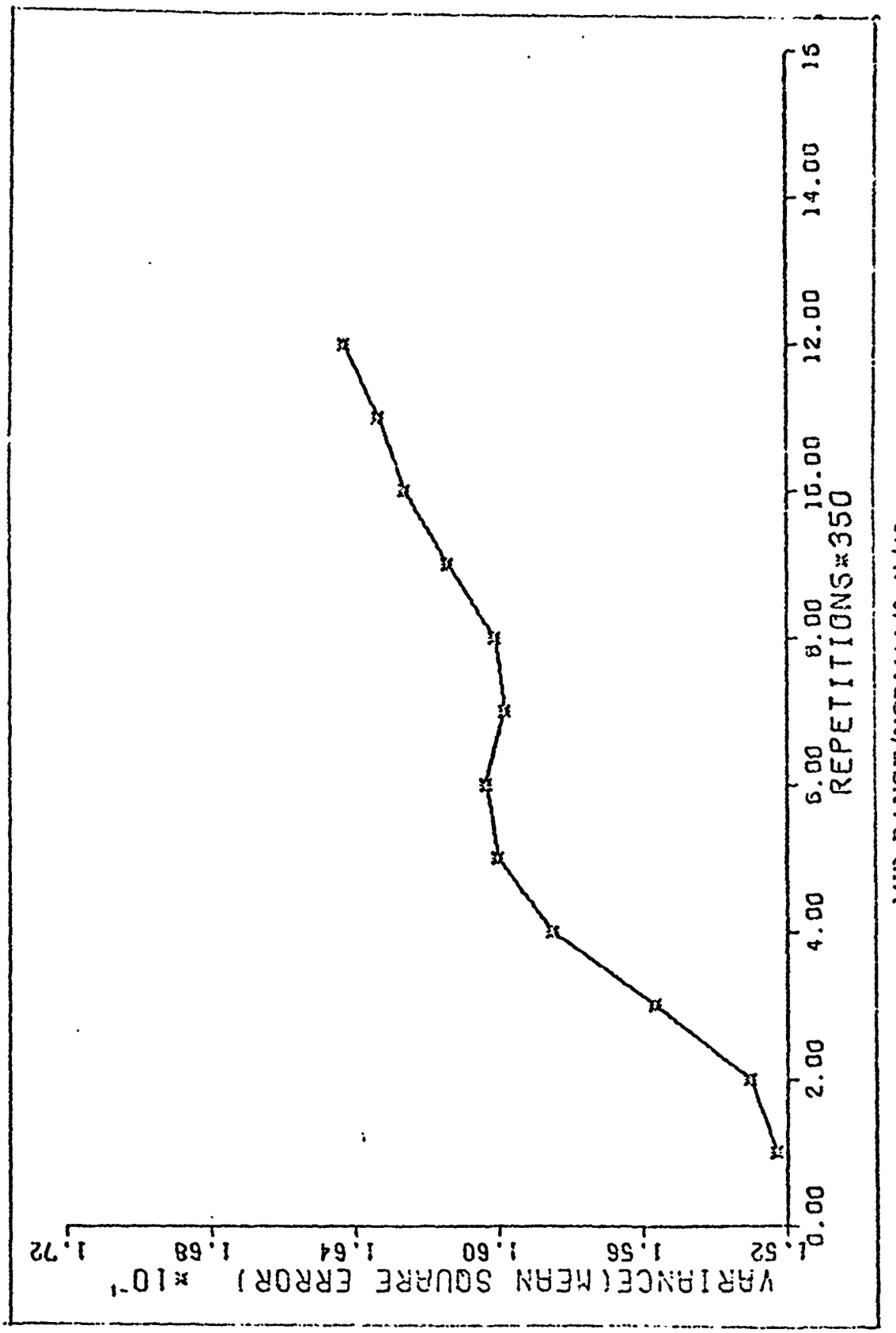
MID RANGE/RECTANGULAR/12



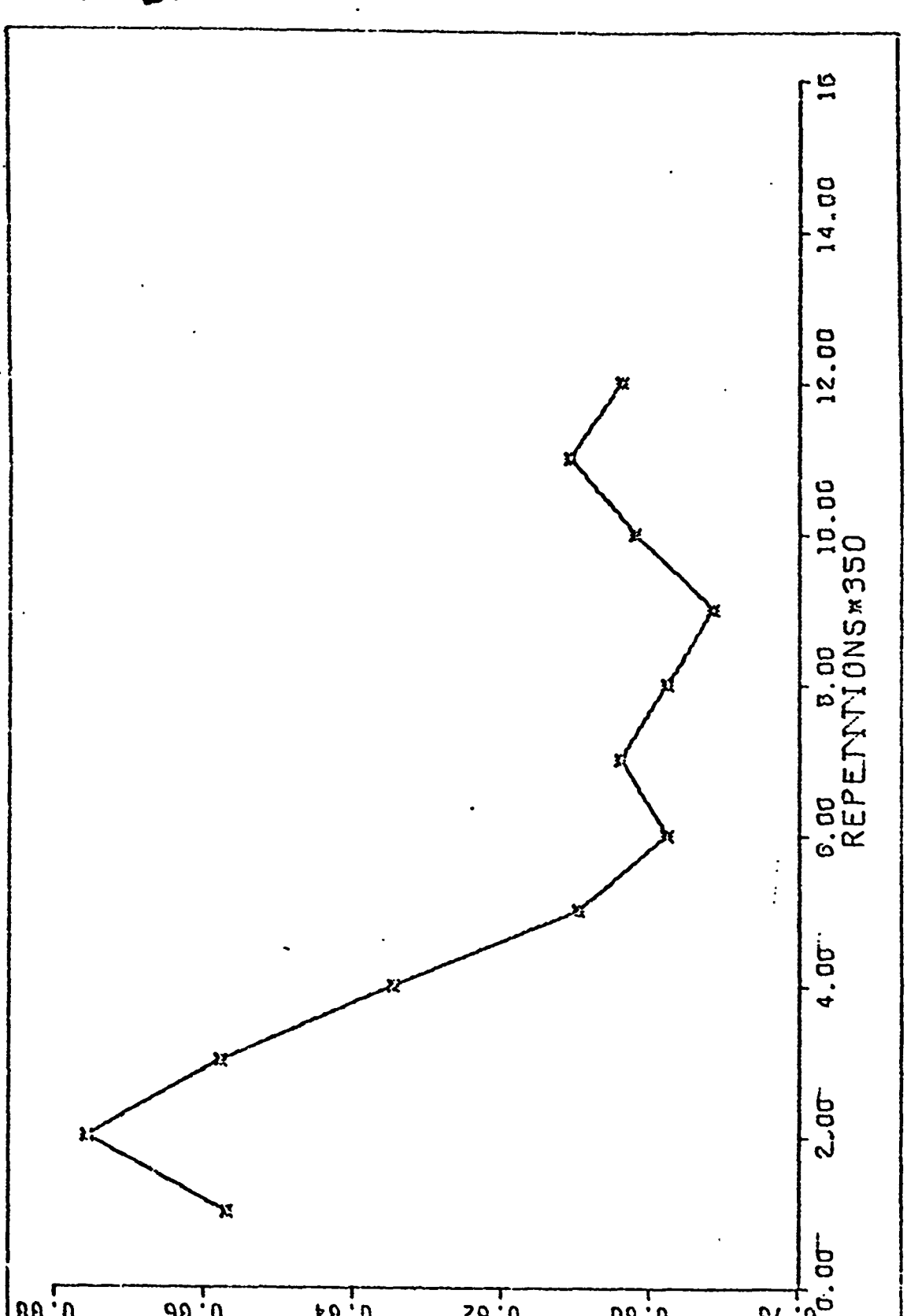
MID RANGE/10% CONTAMINATED NORMAL/12



MID RANGE/TRIANGULAR/12



MID RANGE/NORMAL(0, 1)/12



MID RANGE/DOUB. EXPONENTIAL/12

APPENDIX D

TABLES OF RELATIVE EFFICIENCIES

All efficiencies recorded in the following tables are efficiencies relative to the best estimator consideded for that particular distribution. The best estimator which was used as the base is listed as 100%. In the case of the contaminated normal distribution there are efficiencies greater than 100% recorded. This is because some of the robust estimators actually performed slightly better than the best estimator for that distribution which was the sample mean.

Table 1
Relative Efficiencies For Sample Size 12 *

	Normal (0, 1)	Normal (0, 3)	Doub. Exponential
Hogg's	96.6	96.6	78.0
Hodges-Lehmann	92.1	92.1	94.6
Switzer	77.6	77.6	66.8
Sample Mean	100	100	72.7
Sample Median	67.6	67.6	100
Mid Range	51.0	51.0	14.9

* All efficiencies expressed in per-centage

Table II
Relative Efficiencies For Sample Size 24 *

	Normal (0, 1)	Normal (0, 3)	Doub. Exponential
Hogg's	98.0	98.0	76.4
Hodges-Lehmann	95.0	95.0	88.3
Switzer	78.2	78.2	67.0
Sample Mean	100	100	62.2
Sample Median	67.2	67.2	100
Mid Range	33.4	33.4	06.2

* All efficiencies expressed in per-centage

Table III
Relative Efficiencies For Sample Size 12 *

	Triangular (-1, 1)	Triangular (-5, 5)	Triangular (-10, 10)
Hogg's	97.0	97.0	96.7
Hodges-Lehmann	88.3	88.3	88.3
Switzer	76.9	76.9	76.9
Sample Mean	100	100	100
Sample Median	63.6	63.6	63.6
Mid Range	80.3	81.2	80.3

* All efficiencies expressed in per-centage

Table IV
Relative Efficiencies For Sample Size 24*

	Triangular (-1, 1)	Triangular (-5, 5)	Triangular (-10, 10)
Hogg's	99.1	99.1	99.1
Hodges-Lehmann	88.8	88.8	88.8
Switzer	77.2	77.2	77.2
Sample Mean	100	100	100
Sample Median	63.1	63.1	63.1
Mid Range	78.8	78.8	78.8

* All efficiencies expressed in per-centage

Table V
Relative Efficiencies For Sample Size 12*

	Contaminated 10%	Contaminated 20%	Rectangular
Hogg's	98.5	98.4	46.3
Hodges-Lehmann	98.8	101.4	32.5
Switzer	78.8	80.6	35.5
Sample Mean	100	100	40.7
Sample Median	74.9	79.4	17.5
Mid Range	35.6	32.6	100

* All efficiencies expressed in per-centage

Table VI
Relative Efficiencies For Sample Size 24*

	Contaminated 10%	Contaminated 20%	Rectangular
Hogg's	97.6	97.9	27.3
Hodges-Lehmann	100.1	103.9	18.1
Switzer	80.6	83.5	22.7
Sample Mean	100	100	21.7
Sample Median	73.7	77.3	8.2
Mid Range	19.8	17.8	100

* All efficiencies expressed in per-centage

GSA/MA[redacted]/72-3

VITA

John Caso was born in Philadelphia, Penna., 9 June 1939. He graduated from Monsignor Bonner High School in 1957 and enlisted in the United States Air Force. In 1963 he was selected for the Airmen's Education and Commissioning Program. He subsequently received a Bachelor of Science in Mathematics from Michigan State University and a commission in the USAF in 1965. After completing a radar-electronics course in 1966 he served as an instructor in the course until 1968. He then served a tour as a Radar Maintenance Officer at Indian Mountain Air Force Station, Alaska. Prior to coming to the Air Force Institute of Technology he was Course Supervisor of the OBR3041 Electronics Systems Officer Course at Keesler AFB, Miss.

Permanent Address: 2232 Theresa Ave.
Morton, Penna.